JGR | Machine Learning and Computation

RESEARCH ARTICLE
10.1029/2024JH000322

Modeling Groundwater Levels in California's Central Valley by Hierarchical Gaussian Process and Neural Network Regression

Anshuman Pradhan[1], Kyra H. Adams[2], Venkat Chandrasekaran[1], Zhen Liu[2], John T. Reager[2], Andrew M. Stuart[1], and Michael J. Turmon[2]

[1]Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA, USA, [2]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA

Key Points:

- Groundwater levels in California's Central Valley are sampled irregularly and noisily making groundwater management challenging
- Machine learning methodology of hierarchical Gaussian process and neural network regression is formulated to model groundwater levels
- Modeled uncertainty estimates of 2015–2020 groundwater levels are validated to be consistent with the data distribution of 90 blind wells

Correspondence to:

A. Pradhan,
pradhan1@alumni.stanford.edu

Citation:

Pradhan, A., Adams, K. H., Chandrasekaran, V., Liu, Z., Reager, J. T., Stuart, A. M., & Turmon, M. J. (2024). Modeling groundwater levels in California's Central Valley by hierarchical Gaussian process and neural network regression. *Journal of Geophysical Research: Machine Learning and Computation*, *1*, e2024JH000322. https://doi.org/10.1029/2024JH000322

Received 21 JUL 2024
Accepted 8 OCT 2024

Author Contributions:

Conceptualization: Anshuman Pradhan, Kyra H. Adams, Venkat Chandrasekaran, Zhen Liu, John T. Reager, Andrew M. Stuart, Michael J. Turmon
Data curation: Kyra H. Adams, Zhen Liu, John T. Reager, Michael J. Turmon
Formal analysis: Anshuman Pradhan
Funding acquisition: Venkat Chandrasekaran, Andrew M. Stuart
Methodology: Anshuman Pradhan, Venkat Chandrasekaran, Andrew M. Stuart
Software: Anshuman Pradhan

Abstract Modeling groundwater levels continuously across California's Central Valley (CV) hydrological system is challenging due to low-quality well data which is sparsely and noisily sampled across time and space. The lack of consistent well data makes it difficult to evaluate the impact of 2017 and 2019 wet years on CV groundwater following a severe drought during 2012–2015. A novel machine learning method is formulated for modeling groundwater levels by learning from a 3D lithological texture model of the CV aquifer. The proposed formulation performs multivariate regression by combining Gaussian processes (GP) and deep neural networks (DNN). The hierarchical modeling approach constitutes training the DNN to learn a lithologically informed latent space where non-parametric regression with GP is performed. We demonstrate the efficacy of GP-DNN regression for modeling non-stationary features in the well data with fast and reliable uncertainty quantification, as validated to be statistically consistent with the empirical data distribution from 90 blind wells across CV. We show how the model predictions may be used to supplement hydrological understanding of aquifer responses in basins with irregular well data. Our results indicate that on average the 2017 and 2019 wet years in California were largely ineffective in replenishing the groundwater loss caused during previous drought years.

Plain Language Summary Building a reliable model of groundwater level depths in California's Central Valley (CV) aquifer system is essential for groundwater management and decision-making. However, publicly available water level well data are sparse, irregular, and noisy, resulting in large uncertainties in groundwater modeling efforts. We mathematically formulate a novel machine learning approach, termed the Gaussian process and deep neural network (GP-DNN) regression, to constrain the uncertainties on groundwater levels in CV with input information from a model of the aquifer lithology. The machine learning model uses DNNs to extract useful features from the aquifer lithological model. Subsequently, GP regression approach conducts interpolation in the lithological feature space to predict water levels at every location in the CV along with rigorous estimates of modeling and data uncertainty. The model uncertainty predictions were validated to be reliable by statistically comparing results at 90 wells in the study area that were kept blind during the modeling process. We show how proposed machine learning method may be used to overcome common data limitations in hydrological basins and improve understanding of aquifer response to groundwater recharge and drought recovery.

1. Introduction

With climate change causing frequent periods of intense droughts in many parts of the world, the need for sustainably managing groundwater resources has never been greater. A prime example is California's Central Valley (CV) aquifer system. The CV, supporting a 20 billion dollar per year agricultural industry (Faunt et al., 2016), has been severely strained by recent droughts, depleting groundwater levels and diminishing surface water availability. One of the key variables impacting decisions in groundwater management is understanding how water levels in the underlying aquifer system vary through processes of groundwater recharge and discharge, such as precipitation and agricultural pumping. A key challenge in the CV is the absence of water level measurements that are regularly sampled across space and time. Publicly available groundwater level databases often do not include data from privately owned wells in the valley (D. Johnson & Belitz, 2015; Kim et al., 2021). Available measurements are often noisy and sampled with large temporal gaps (see Section 3 for additional discussion). These factors impede developing a comprehensive understanding of how CV hydrogeology and

Supervision: Venkat Chandrasekaran, Andrew M. Stuart
Validation: Anshuman Pradhan, Kyra H. Adams, Zhen Liu, John T. Reager, Michael J. Turmon
Visualization: Anshuman Pradhan
Writing – original draft: Anshuman Pradhan
Writing – review & editing: Anshuman Pradhan, Kyra H. Adams, Venkat Chandrasekaran, Zhen Liu, John T. Reager, Andrew M. Stuart, Michael J. Turmon

groundwater processes such as recharge, depletion, and aquifer subsidence are interlinked, rendering decision-making for water resources management difficult.

A traditional approach to comprehensively understanding subsurface water levels is to model the groundwater level or hydraulic head of the aquifer as the response of a physics-based groundwater flow model (Faunt 2009; Harbaugh, 2005; Todd & Mays, 2005). 3D simulation of groundwater flow requires a conceptual model of the aquifer and quantification of rock hydraulic and storage properties. Since subsurface properties are not observed exhaustively, it becomes necessary to derive them inversely to be able to perform flow simulation. In addition to head measurements at wells, remote sensing (Chaussard et al., 2014, Z. Liu et al., 2019) and hydrogeophysical data may be used in the inversion process (Binley et al., 2015; Kang et al., 2021; Smith & Knight, 2019). Given the sparse and noisy nature of available water data in CV, it is highly desirable to quantify any uncertainty associated with the modeling process. If the prior estimates of uncertainty are modeled by probability distributions, the Bayesian inversion paradigm may then be employed to condition the prior uncertainty to observed data (Stuart, 2010; Tarantola, 2005). A large source of uncertainty for subsurface property inversion relates to the model parameterization (Caers, 2011), for instance specifying the structure of faults and the stratigraphy of aquifer confining units, as well as the spatial distribution of lithofacies and intra-facies aquifer property variability.

Over the past few decades, research developments in the geostatistics literature have led to the development of several sophisticated models for spatial heterogeneity associated with subsurface reservoirs, including covariance based, training image based, object-based, surface-based and process-mimicking models (Caers, 2005; Deutsch & Journel, 1998; Mariethoz & Caers, 2014; Pyrcz & Deutsch, 2014). Covariance-based models include variants of kriging or Gaussian process regression method, which enforce constraints on the two-point correlations or second order statistics in the modeling domain. Constraining just the spatial covariance model has been often found to be limited in replicating the complex geological heterogeneity associated with common geological settings, for instance a fluvial aquifer system containing channelized sand lithofacies (Feyen & Caers, 2006; Linde et al., 2015). To address these challenges, several methods for modeling higher-order spatial correlations have been developed, a few of which were listed previously. However, such models often rely on techniques from computer-vision to stitch together complex spatial patterns or drop geological objects into a modeling grid, and are not amenable for conditioning to dense geophysical or remote-sensing data (Bertoncello, 2011; Pradhan, 2020), unlike kriging-based approaches. Stochastic search methods, such as Markov Chain Monte-Carlo (MCMC), are often necessitated for model inversion with uncertainty quantification (Hermans et al., 2015; Keating et al., 2013; Laloy et al., 2013; Mariethoz et al., 2010). Stochastic search methods are known for their computational complexity, especially when optimizing for 3D aquifer models with millions of grid cells.

Given the several limitations associated with subsurface modeling, we adopt a data-driven approach to groundwater modeling and uncertainty quantification. In many cases, the final groundwater management decisions are not directly dependent on the earth model itself, but rather on some summary statistics or trends of groundwater variability. For instance, water level long-term and seasonal variability trends may be used to understand aquifer response to groundwater recharge and discharge (Neely et al., 2021; Riel et al., 2018), aiding groundwater budgeting efforts during dry years. In this paper, we seek to estimate these trends continuously across CV using sparse well data. We propose a novel methodology combining Gaussian process (GP) regression and deep neural networks (DNN) to achieve this objective. Gaussian process regression, also known as kriging, was introduced almost 70 years ago by Krige (1951), with subsequent theoretical development made by Matheron (1962). Since then, kriging and its several variants have seen widespread usage in geosciences and spatial statistics (Cressie, 1993; Goovaerts, 1997; Journel & Huijbregts, 1978). Our primary motivation for employing the GP regression methodology is that it allows fast and easy derivation of uncertainty estimates. However, a fundamental limitation of kriging is its assumption of spatial stationarity across the modeling domain. This limits the ability to model non-stationary data, for instance data with varying spatial length-scales across different regions.

In geostatistics literature, non-stationarity has been handled with techniques such as kriging with locally varying mean, universal kriging and intrinsic random functions (de Marsily, 1987). Kriging has received emerging interest in the machine learning literature where it is more commonly known as GP regression (Rasmussen & Williams, 2006). This has led to the development of several sophisticated mathematical formulations for handling non-stationary data with GPs. Two broad categories of developments may be identified. In the first category, the

non-stationarity challenge is addressed by construction of hierarchical formulations of covariance kernels. Paciorek (2003) propose to model non-stationary data by GPs whose covariance function depends on another GP. Damianou and Lawrence (2013) proposed the methodology known as deep Gaussian processes, where each GP layer with stationary covariance, is composed from another stationary GP. Dunlop et al. (2018) extend Paciorek (2003)'s work on multiple hierarchical layers of covariance functions and propose to handle non-stationarity by iteratively modifying the length-scales of covariance functions. In Roininen et al. (2019)'s paper, the hierarchy is built using the stochastic partial differential equation representation of GP (Lindgren et al., 2011). While these complex kernels have some shown promising results, they lack the representational capabilities in high dimensions that has become commonplace in machine learning with DNNs (Bradshaw et al., 2017). The other category involves applying standard GP regression in a latent space obtained by deformation of the input feature space such that the assumptions of stationarity hold in this latent space. P. D. Sampson and Guttorp (1992) proposed obtaining the latent space by warping the geographical coordinate space. Specifically, multidimensional scaling (MDS) of the training data was conducted such that empirical estimates of the variogram were preserved during MDS, and subsequently spline mappings were used to derive a smooth latent field in the MDS space. In contrast to two step modeling of the latent space, Schmidt and O'Hagan (2003) proposed a fully Bayesian approach that involves performing the spatial deformation by Gaussian process priors. However, their approach requires expensive MCMC algorithms to sample the posterior. P. Sampson et al. (2001) provide a review of some of the original works involving warping of the input space. Recently, there has been a resurgence of interest in this approach, especially on modeling the latent space by DNNs. Calandra et al. (2016) and Wilson et al. (2016) perform GP regression in the latent space learned with neural networks, referring to their approaches as manifold GP regression and deep kernel learning respectively. Bradshaw et al. (2017) apply the GP hybrid DNN model to image classification task. It was demonstrated how deep neural networks boosted the capability of GP to model non-stationary, discontinuous and noisy data. Note that the above approaches were formulated in the univariate regression or classification setting.

In this paper, we extend the above formulation to regression in the multivariate setting with geospatially and hydrogeologically indexed data. By hydrogeological indexing, we refer to the lithological data that were also used as input features to the model in addition to geospatial coordinates. We establish a two-level model hierarchy with a DNN below a GP layer, trained end-to-end. As shown in Section 3, proposed model is able to handle the non-stationary, sparse and noisy well data by learning to appropriately scale the coordinates of the latent space. The model allows analytical derivation of the posterior uncertainty and fast generation of posterior samples. The intended novel contributions of this paper are as follows.

1. We extend the formulation of manifold GP regression or deep kernel learning to handle geospatially and hydrogeologically indexed data with multivariate prediction variables. The novel methodology is simply referred to as GP-DNN regression.
2. We formulate a novel cross-validation technique using chi-square quantile-quantile plots for evaluating generalization capability of GP-based regression models.
3. We present a real world case study of groundwater level modeling in CV by GP-DNN regression. Specifically, we assume a linear model for seasonal and long-term variability of groundwater levels and demonstrate how the machine learning methodology may be used with this groundwater model, and irregular data sets to yield statistically valid uncertainty estimates without detailed physics-based modeling.

A theoretical description of the proposed methodology, followed by the real world case study from California's CV is presented next. The application is focused around modeling long-term and seasonal trends in groundwater levels in CV from 2015 to 2020. We also provide interpretations and visualizations of the latent space learned by the DNN, explicitly demonstrating how it handles non-stationarity and uncertainty. We present hydrological interpretations of the seasonal and long-term trends of water levels in CV, especially in the context of drought and recovery to illustrate how this method may be applied to real-life understanding of hydrologic data. Finally, we discuss future work, both in terms of how the methodology maybe extended to handle more complicated real-world data noise scenarios as well how the CV groundwater model may be made more rigorous by employing the methods presented in this paper.

2. Methodology

Let $\mathcal{D} \in \mathbb{R}^2$ denote the 2D geospatial domain of the area of interest. Let $\boldsymbol{x} \in \mathcal{D}$ denote the vector for spatial coordinates, t denote time and u denote the true water level depths. Noisy observations of water levels, $\{(\boldsymbol{x}_i, t_j, u_{obs}(\boldsymbol{x}_i, t_j)); i \in \{1, \ldots, m\}, j \in \{1, \ldots, n_i\}\}$ are available at m discrete water well locations, each with an irregular number of n_i samples across time. In this paper, the variables of interest are long-term and seasonal trends of water level fluctuations. In Section 3, we model water levels $u(\boldsymbol{x}, t)$ at any $\boldsymbol{x}$ by combining linear and sinusoidal time-series models quantifying the long-term and seasonal water level fluctuations respectively. Note that we make the simplifying assumption that water levels vary only across 2D space $\boldsymbol{x}$ as discussed in Sections 3.2.1 and 3.2.2. The prediction variables consist of the temporal model parameters, that is, intercept and slope parameters of the linear model, and amplitude and phase parameters of the sinusoidal model at all $\boldsymbol{x} \in \mathcal{D}$. The preceding four temporal model parameters at each spatial location are denoted as $\boldsymbol{y} \in \mathbb{R}^d; d = 4$. The formulation we present below thus applies to multivariate prediction variables. Given observations of u at wells, noisy estimates $\{\boldsymbol{y}(\boldsymbol{x}_i); i = 1, \ldots, m\}$ of the prediction variables may be obtained by linear regression at well locations as discussed in Section 3.2.1. The goal is to recover the true underlying signal of the prediction variables at any $\boldsymbol{x}$.

The proposed approach of this paper is to model the prediction variables as multivariate Gaussian random processes. A random process is defined as an indexed collection of random vectors. A specific advantage of the GP formulation is that given a set of irregularly sampled observations, the posterior predictive distribution at any query index may be derived analytically. In many cases, prediction variables $\boldsymbol{y}$ wont be normally distributed. A normal score histogram transformation (see Section 3.2.1) may be applied in that case and the GP formulation is specified on the normally transformed variables.

2.1. Spatial GP Regression and Limitations

Consider the baseline case when the GP $a(\boldsymbol{x}) : \mathbb{R}^2 \to \mathbb{R}^d$ is indexed over geospatial coordinates $\boldsymbol{x}$. Specifically,

$$a(\boldsymbol{x}) \sim \mathcal{GP}\left(0, k\left(\boldsymbol{x}, \boldsymbol{x}'\right)\right), \tag{1}$$

with zero mean and covariance kernel $k : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^{d \times d}$. The covariance kernel $k\left(\boldsymbol{x}, \boldsymbol{x}'\right)$ encapsulates the prior assumptions on the spatial heterogeneity of the random process by quantifying the similarity between two input locations $\boldsymbol{x}$ and $\boldsymbol{x}'$. Given the above a-priori Gaussian assumptions and well observations, the posterior predictive distribution at any query location $\boldsymbol{x}_*$ may be derived by multivariate Gaussian process (GP) regression, also known as cokriging. Cokriging is a very mature methodology for spatial interpolation and regression (see references in Section 1). Several variants of the basic kriging approach have been proposed such as simple cokriging, ordinary cokriging and universal cokriging distinguished by how the mean of the GP is specified. In this paper, we work with the simple cokriging approach in which the mean is specified to be a constant across $\mathcal{D}$ as shown in Equation 1.

The main limitations associated with the cokriging approach are as follows:

1. In typical geological settings, repeated measurements of prediction variables at two distinct locations $\left\{\boldsymbol{y}(\boldsymbol{x}), \boldsymbol{y}\left(\boldsymbol{x}'\right)\right\}$ are seldom available to make inferences about the form of $k\left(\boldsymbol{x}, \boldsymbol{x}'\right)$. Therefore, a decision of spatial stationarity over $\mathcal{D}$ is made by assuming $k\left(\boldsymbol{x}, \boldsymbol{x}'\right)$ depends only on the distances between the input locations as $k\left(\left\|\boldsymbol{x} - \boldsymbol{x}'\right\|_2\right)$, where, $\|.\|_2$ denotes ℓ_2 norm (Goovaerts, 1997). Note that under the above assumptions, the covariance kernel will be invariant to spatial translations. Thus, any two locations separated by identical distances will be assigned identical covariances. Since hydrological and hydrogeological data commonly exhibit significantly varying spatial correlation scales across the hydrological basin, stationary kernels may lead to overly smooth or rough processes, limiting the predictive ability of the model.
2. As defined later, standard covariance kernels model a range of influence through their length-scale parameters. In other words, posterior predictive uncertainty at a test location beyond the range of influence of any well location will be large. Thus, regression over $\mathcal{D}$ will lead to large posterior predictive uncertainty if the observed data is sparsely sampled, which is the case with CV well data.

A potential solution is to perform the GP regression in an extended feature space $\hat{\boldsymbol{x}} = \left[\boldsymbol{x}, \check{\boldsymbol{x}}\right]^T, \check{\boldsymbol{x}} \in \mathbb{R}^n$, with T being the transpose operator. For modeling groundwater levels, $\check{\boldsymbol{x}}$ could pertain to the aquifer hydrogeology, for

example, the depth and lithologic composition of aquifer stratigraphies (see Section 3.2.2 for context within the CV hydrological basin). The requirement is that $\check{\boldsymbol{x}}$ must be known for every $\boldsymbol{x} \in \mathcal{D}$ such that random process may be indexed in the combined geospatial and hydrogeological space $\hat{\boldsymbol{x}}$ as $a(\hat{\boldsymbol{x}})$. Indexing $a(.)$ over $\hat{\boldsymbol{x}}$ will address the training data sparsity limitation when testing coordinates have more similarity to the training data coordinates in $\hat{\boldsymbol{x}}$-space versus $\boldsymbol{x}$-space. However, note that this does not explicitly bypass the stationarity assumption since the true $a(\hat{\boldsymbol{x}})$ may also exhibit non-stationary behavior over $\hat{\boldsymbol{x}}$. To handle this, we propose using the GP-DNN formulation which learns to appropriately re-configure the distances between the random process index.

2.2. Multivariate Regression by Hierarchical GP-DNN

As motivated in Section 1, several authors have proposed to handle non-stationary data by modeling a latent space that is able to accommodate the assumptions of stationarity,

$$\tilde{\boldsymbol{x}} = \phi(\hat{\boldsymbol{x}}; \boldsymbol{\theta}), \tag{2}$$

where $\tilde{\boldsymbol{x}} \in \mathbb{R}^p$ denotes the latent feature space and $\phi : \mathbb{R}^{n+2} \to \mathbb{R}^p$ is a feature projection map to be inferred by machine learning. The dimensionality of input feature space is $n + 2$ since it also includes the 2-dimensional geospatial coordinates. In the proposed GP-DNN hierarchical model, $\phi(.;\boldsymbol{\theta})$ is taken to be a deep neural network with learnable parameters $\boldsymbol{\theta}$. The Gaussian random process predictive prior will be indexed over the latent feature space,

$$a(\tilde{\boldsymbol{x}}) \sim \mathcal{GP}\big(0,\, k(\tilde{\boldsymbol{x}},\, \tilde{\boldsymbol{x}}')\big), \tag{3}$$

where $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^{d\times d}$ is the covariance kernel. Note that the usage of the term *predictive* indicates probability distributions defined over the prediction variables. We show next how the GP prior predictive distribution may be conditioned to data observations to derive the posterior predictive distribution.

2.2.1. The Generative Model and Conditioning to Training Data

In the following, subscripts τ and $*$ are used to denote training and test data for machine learning. Let $\boldsymbol{y}_\tau$ denote the $dm \times 1$ vector constituting noisy observations of the true signal $a(\tilde{\boldsymbol{x}})$ corresponding to well locations and $\tilde{X}_\tau$ is the $dm \times p$ latent feature matrix, which is computed from the original $dm \times (n + 2)$ input feature matrix $\hat{X}_\tau$ using the DNN. d is the dimensionality of the prediction variable and m is the number of training samples. Since $a(\tilde{\boldsymbol{x}})$ is assumed to be a GP, its' realizations corresponding to the well locations, denoted by the vector $\boldsymbol{a}_\tau$, will be distributed according to a multivariate Gaussian distribution. We assume that realizations of $\boldsymbol{a}_\tau$ are subsequently corrupted by Gaussian noise yielding the noisy measurements $\boldsymbol{y}_\tau$. To summarize, the generative model for the well data is specified as,

$$\tilde{X}_\tau = \phi(\hat{X}_\tau),\;\; \boldsymbol{a}_\tau|\tilde{X}_\tau \sim \mathcal{N}\big(\boldsymbol{0},\, \tilde{K}_{\tau\tau}\big),\;\; \text{and } \boldsymbol{y}_\tau|\boldsymbol{a}_\tau \sim \mathcal{N}(\boldsymbol{a}_\tau,\, \Sigma_n), \tag{4}$$

where $\mathcal{N}(.,.)$ denotes the multivariate Gaussian distribution and $\tilde{K}_{\tau\tau}$ is the $dm \times dm$ covariance matrix (see Section 2.2.2 for details). Note that operator $\phi(.;\boldsymbol{\theta})$, in the first equality above, operates on each row of matrix $\hat{X}_\tau$ and we have abused notation for brevity. The covariance matrix for the noise distribution is specified as $\Sigma_n = \text{diag}\big([\boldsymbol{\sigma}_{n_1},\, \ldots,\, \boldsymbol{\sigma}_{n_d}]^T\big)$, where $\boldsymbol{\sigma}_{n_i} = \big[\sigma^2_{n_i},\, \ldots,\, \sigma^2_{n_i}\big]^T$ is a $m \times 1$ vector specifying identical observational noise variance $\sigma^2_{n_i}$ for each prediction variable $a(\boldsymbol{x})_i$. The covariance kernel parameters, noise levels and DNN architectural variables are treated as hyper-parameters of the model, to be tuned by cross-validation as described in Section 3.3 (see Table C2 for a complete list of hyper-parameters). Parameters $\boldsymbol{\theta}$ of the DNN are trained as discussed in Section 2.2.3.

Let $\boldsymbol{a}_*$ denote the realization of $a(\hat{\boldsymbol{x}})$ at test location $\hat{\boldsymbol{x}}_*$. The posterior predictive distribution of $\boldsymbol{a}_*$ is obtained by conditioning the prior predictive distribution to the noisy observations. The joint prior distribution of $\boldsymbol{y}_\tau$ and $\boldsymbol{a}_*$, given the training and test locations, may be specified as

$$\begin{bmatrix} \boldsymbol{y}_\tau \\ \boldsymbol{a}_* \end{bmatrix} | \hat{X}_\tau, \hat{\boldsymbol{x}}_* \sim \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix} \tilde{K}_{\tau\tau} + \Sigma_n & \tilde{K}_{\tau *} \\ \tilde{K}_{*\tau} & \tilde{K}_{**} \end{bmatrix} \right) \tag{5}$$

(Rasmussen & Williams, 2006), where, $\tilde{K}_{\tau\tau}$ is the covariance matrix for the training locations, $\tilde{K}_{*\tau}$ and $\tilde{K}_{\tau *}$ contain the covariances between the training and test locations, while $\tilde{K}_{**}$ is the covariance matrix for test locations. Since the joint distribution of $\boldsymbol{y}_\tau$ and $\boldsymbol{a}_*$ is a Gaussian distribution, the distribution of $\boldsymbol{a}_*$, conditioned on the training observations and the training and test features, is also a Gaussian distribution derived as

$$\boldsymbol{a}_* | \boldsymbol{y}_\tau, \hat{X}_\tau, \hat{\boldsymbol{x}}_* \sim \mathcal{N}\left(\tilde{K}_{*\tau}\left[\tilde{K}_{\tau\tau} + \Sigma_n\right]^{-1} \boldsymbol{y}_\tau, \tilde{K}_{**} - \tilde{K}_{*\tau}\left[\tilde{K}_{\tau\tau} + \Sigma_n\right]^{-1} \tilde{K}_{\tau *} \right). \tag{6}$$

Here, $\tilde{K}_{*\tau}\left[\tilde{K}_{\tau\tau} + \Sigma_n\right]^{-1} \boldsymbol{y}_\tau$ is the cokriging estimate of the posterior predictive mean and $\tilde{K}_{**} - \tilde{K}_{*\tau}\left[\tilde{K}_{\tau\tau} + \Sigma_n\right]^{-1} \tilde{K}_{\tau *}$ is the posterior predictive covariance.

2.2.2. The Multivariate Kernel

The assumption of covariance kernel stationarity is made in the latent space by taking

$$k\left(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{x}}'\right) = k\left(\left\| \tilde{\boldsymbol{x}} - \tilde{\boldsymbol{x}}' \right\|_2 \right), \ \forall \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{x}}'. \tag{7}$$

By definition, the covariance kernel should be a symmetric, positive definite function (Paciorek, 2003). Several valid covariance functions have been studied in the geostatistical and ML literature such as the squared exponential and the Matérn kernel. To ensure kernel validity in the multivariate regression setting, we employ the linear model of coregionalization (De Iaco et al., 2003; Journel & Huijbregts, 1978) which models all the components of the multivariate process as linear combinations of the same underlying permissible random processes. This states that the kernel, when specified as

$$k\left(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{x}}'\right) = \sum_i K^i_{amp} k^i_{valid}\left(\left\| \tilde{\boldsymbol{x}} - \tilde{\boldsymbol{x}}' \right\|_2 \right), \tag{8}$$

will be a valid positive semi-definite kernel if K^i_{amp} is a $d \times d$ positive semi-definite matrix and $k^i_{valid} : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is a permissible positive semi-definite kernel for each i. K^i_{amp} contains the variance and covariance scaling factors for the prediction variables. If each component of $a(.)$ is normalized to have unit variance, then the diagonal elements of K^i_{amp} will have unit magnitude and off-diagonal elements specify the correlation coefficient for each variable pair. In this paper, we choose $i = 1$ and k_{valid} to be the Matérn kernel $k^{\nu=2.5}_{Matérn} : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$, where ν is the Matérn parameter controlling the roughness of the kernel. The Matérn kernel allows greater control on the roughness of the random process through the ν parameter (Stein, 1999) and $\nu = 2.5$ is a common choice for machine learning applications (Rasmussen & Williams, 2006). Appendix A contains additional details on how the covariance matrices are constructed from the Matérn kernel.

2.2.3. GP-DNN End-To-End Training

DNN parameters $\boldsymbol{\theta}$ will be trained end-to-end along with the GP layer by maximization of the data likelihood distribution as described below. The dimensionality p of the latent feature space $\tilde{\boldsymbol{x}}$ and the DNN architectural parameters, such as the number of hidden layers and the number of neurons in each hidden layer are treated as hyper-parameters, to be tuned by cross-validation. From Equation 4, it follows that the training data likelihood

$$\boldsymbol{y}_\tau | \hat{X}_\tau; \boldsymbol{\theta} \sim \mathcal{N}\left(\boldsymbol{0}, \tilde{K}_{\tau\tau} + \Sigma_n \right). \tag{9}$$

Since the Gaussian likelihood has an analytical expression, parameters $\boldsymbol{\theta}$ may be estimated by minimizing the negative of log-likelihood

$$\log p(\boldsymbol{y}_\tau|\hat{X}_\tau;\boldsymbol{\theta}) = -\frac{1}{2}\boldsymbol{y}_\tau^T\left[\tilde{K}_{\tau\tau}+\Sigma_n\right]^{-1}\boldsymbol{y}_\tau - \frac{1}{2}\log|\tilde{K}_{\tau\tau}+\Sigma_n| + \text{constant} \tag{10}$$

(Bradshaw et al., 2017; Calandra et al., 2016; Rasmussen & Williams, 2006; Wilson et al., 2016). Taking derivatives of $\log p(\boldsymbol{y}_\tau|\hat{X}_\tau)$ w.r.t the parameter θ_k, we obtain

$$\begin{aligned}\frac{\partial}{\partial\theta_k}p(\boldsymbol{y}_\tau|\hat{X}_\tau;\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{y}_\tau^T\left[\tilde{K}_{\tau\tau}+\Sigma_n\right]^{-1}\frac{\partial\tilde{K}_{\tau\tau}}{\partial\theta_k}\left[\tilde{K}_{\tau\tau}+\Sigma_n\right]^{-1}\boldsymbol{y}_\tau \\ -\frac{1}{2}tr\left(\left[\tilde{K}_{\tau\tau}+\Sigma_n\right]^{-1}\frac{\partial\tilde{K}_{\tau\tau}}{\partial\theta_k}\right).\end{aligned} \tag{11}$$

In the equality above, we used identities for derivative of inverse matrix, $\frac{\partial K^{-1}}{\partial\theta} = -K^{-1}\frac{\partial K}{\partial\theta}K^{-1}$, and the derivative of matrix log determinant $\frac{\partial\log|K|}{\partial\theta} = tr\left(K^{-1}\frac{\partial K}{\partial\theta}\right)$. Given that the $k^{\nu=2.5}_{Matérn}(.)$ is a differentiable function w.r.t to its inputs, the entries of $\frac{\partial\tilde{K}_{\tau\tau}}{\partial\theta_k}$ may be analytically derived by the back-propagation algorithm Bishop (2006). Parameters $\boldsymbol{\theta}$ may then be trained by typical stochastic gradient descent algorithms commonly employed to train deep learning networks (Algorithm 1). Additional details on training and hyper-parameter tuning are provided in Section C2 in Appendix C.

Algorithm 1. Algorithm for Training of the GP-DNN Hierarchical Model.

Input:
Training data $\{\hat{X}_\tau, \boldsymbol{y}_\tau\}$ and validation data $\{\hat{X}_*, \boldsymbol{y}_*\}$
Hyper-parameters listed in Table C2, length-scales $\boldsymbol{l} = \mathbf{1}$
Parameters $\boldsymbol{\theta}$ initialized using the Glorot uniform distribution (Glorot & Bengio [2010])
n_{epoch}: number of training epochs.

for $i = 1, \ldots, n_{epoch}$ **do**
- Compute $\tilde{\boldsymbol{x}}_{\tau_i} = \phi(\hat{\boldsymbol{x}}_{\tau_i};\boldsymbol{\theta})\,\forall\, i \in \{1,\ldots,m\}$, and $\tilde{X}_\tau$.
- Formulate $\tilde{K}_{\tau\tau}$ as given by equation A1.
- Compute the negative log-likelihood of the GP marginal distribution $\log p(\boldsymbol{y}_\tau|\hat{X}_\tau;\boldsymbol{\theta})$ as given in equation 10 and save to training history.
- Derive $\frac{\partial}{\partial\theta_k}p(\boldsymbol{y}_\tau|\hat{X}_\tau;\boldsymbol{\theta}), \forall\, k$ using equation 11, where $\frac{\partial\tilde{K}_{\tau\tau}}{\partial\theta_k}$ is calculated by the back-propagation algorithm.
- Use the Adam stochastic optimization algorithm (Kingma & Ba [2015]) to update $\boldsymbol{\theta}$ and save to training history.
- Estimate the mean and covariance matrix of the posterior predictive distribution $\boldsymbol{a}_*|\boldsymbol{y}_\tau, \hat{X}_\tau, \hat{X}_*$ (similar to equation 6).
- Evaluate the cross-validation statistic according to equation 12 and save to training history.

end

Output:
Trained $\boldsymbol{\theta}$: set to that estimated after epoch with highest cross-validation metric
Training history: evolution of likelihood of training data, cross-validation metric and $\boldsymbol{\theta}$ across all epochs.

2.2.4. Cross-Validation Statistics for GP Based Regression

The generalization power of GP regression models will be assessed with a test set constituting of randomly sampled well locations that will be held out and kept blind during model training and hyper-parameter tuning. These wells will be referred to as blind wells in our study. Following is a discussion of two specific cross-validation statistics evaluated on the test set, (a) likelihood under the posterior predictive distribution, and (b) deviation of the chi-square plot from the identity function, that are used to compare the GP regression model performance in Section 3.3.

2.2.4.1. Posterior Predictive Likelihood

Let $\hat{X}_*$ be the $dm_* \times (n+2)$ be the feature matrix and $\boldsymbol{y}_*$ be the $dm_* \times 1$ vector containing observations of the prediction variables computed from the test set. Following from Equation 6, the negative log-likelihood of the Gaussian posterior predictive distribution on test set prediction variables given the training set and test feature matrix

$$-\log p(\boldsymbol{y}_*|\hat{X}_\tau, \boldsymbol{y}_\tau, \hat{X}_*) = \frac{1}{2}[\boldsymbol{y}_* - \boldsymbol{\mu}]^T K^{-1}[\boldsymbol{y}_* - \boldsymbol{\mu}] + \frac{1}{2}\log|K| + \frac{dm_*}{2}\log 2\pi, \tag{12}$$

where, $\boldsymbol{\mu} = \tilde{K}_{*\tau}\left[\tilde{K}_{\tau\tau} + \Sigma_n\right]^{-1}\boldsymbol{y}_\tau$ and $K = \tilde{K}_{**} - \tilde{K}_{*\tau}\left[\tilde{K}_{\tau\tau} + \Sigma_n\right]^{-1}\tilde{K}_{\tau*}$ are the cokriging estimates of the posterior predictive mean and covariance. The first term in the R.H.S. of Equation 12 may be interpreted as half of the squared Mahalanobis distance (Mahalanobis, 1936)

$$\mathcal{M}_D^2(\boldsymbol{y}_*; \boldsymbol{\mu}, K) = [\boldsymbol{y}_* - \boldsymbol{\mu}]^T K^{-1}[\boldsymbol{y}_* - \boldsymbol{\mu}]. \tag{13}$$

Given mean $\boldsymbol{\mu}$ and covariance matrix K, Mahalanobis distance computes the statistically standardized distance of $\boldsymbol{y}_*$ from the mean. The inverse covariance matrix K^{-1} serves the purpose of standardizing each coordinate of $\mathbb{R}^{dm_*}$ by corresponding variance and removing inter-coordinate correlations as estimated by the posterior predictive distribution (Etherington, 2019). The Mahalanobis distance thus acts as an indicator of the fit of the test data under the estimated posterior predictive distribution. The second and third terms in the R.H.S. of Equation 12 relate to the normalization constant of the multivariate Gaussian probability density and consequently the volume under the multivariate density function. For a given dimensionality of the output space, $\frac{1}{2}\log|K|$ could be interpreted in terms of the predictive model complexity. $|K|$ is the generalization of variance in multivariate settings (Wilks, 1932), and thus a larger value for $|K|$ implies a more complex model that will be able to explain larger variability in the data. In other words, rewarding lower values of the negative log-likelihood statistic encourages simpler models that fit the data better.

2.2.4.2. Chi-Square Quantile-Quantile Plots for GP

Going beyond the data likelihood, we propose to employ quantile-quantile (Q-Q) plots to assess the goodness of fit of the predicted uncertainty intervals to the test data. The Q-Q plot is a standard statistical tool for evaluating whether empirical data belong to a specified theoretical probability distribution through a graphical comparison of the empirical quantiles to their theoretical counterparts (Gnanadesikan & Wilk, 1968). While it is generally difficult to generalize Q-Q plots to multivariate data (Easton & McCulloch, 1990), multivariate normality can be tested using the chi-square plot (R. A. Johnson & Wichern, 2007, see Section 4.6). We show how the chi-square plot approach may be extended to multivariate GPs. For notational simplicity, we present the results considering spatial GP regression but it is straightforward to extend the treatment to GP-DNN. Consider a randomly selected set of feature coordinates $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{m_*}\}$ from $\mathcal{D}$ where noisy observations $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{m_*}\}$ of d-variate prediction variables are available. The objective is to determine the accuracy of the hypothesis that each observation $\boldsymbol{y}_i$ is a sample from the corresponding predictive posterior distribution in the set

$$\{\mathcal{N}(\boldsymbol{\mu}_1, K_1), \mathcal{N}(\boldsymbol{\mu}_2, K_2), \ldots, \mathcal{N}(\boldsymbol{\mu}_{m*}, K_{m*})\},$$

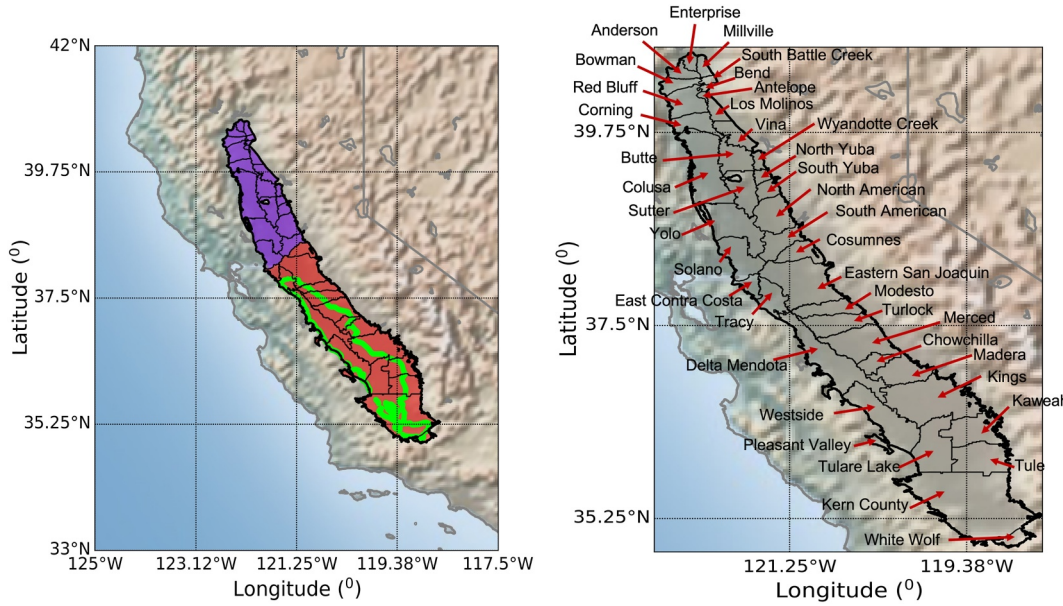

Figure 1. Extent of CV outlined in black with groundwater subbasins. (left) Sacramento and San Joaqin Valley areas have been colored by purple and red. Mapped extent of the Corcoran clay is shown in lime green. (right) Nomenclature of the CV's groundwater subbasins.

where $\mathcal{N}(\boldsymbol{\mu}_i, K_i)$ is distribution of $a(\boldsymbol{x}_i)|X_\tau, \boldsymbol{y}_\tau, \boldsymbol{x}_i$ as estimated using Equation 6 $\forall \boldsymbol{x}_i \in \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{m_*}\}$. As proved by R. A. Johnson and Wichern (2007, see Result 4.7),

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, K) \Rightarrow \mathcal{M}_D^2(\boldsymbol{y};\boldsymbol{\mu}, K) \sim \mathcal{X}_d^2, \tag{14}$$

where $\mathcal{X}_d^2$ is the chi-square distribution having d degrees of freedom, with $\mathcal{M}_D^2(.)$ being estimated by Equation 13. It directly follows from this result that if the test locations were sampled independently and the GP regression robustly estimated the associated posterior predictive means and covariance matrices, the set

$$\{\mathcal{M}_D^2(\boldsymbol{y}_1;\boldsymbol{\mu}_1, K_1)\}, \mathcal{M}_D^2(\boldsymbol{y}_2;\boldsymbol{\mu}_2, K_2), \ldots, \mathcal{M}_D^2(\boldsymbol{y}_{m_*};\boldsymbol{\mu}_{m_*}, K_{m_*})\}$$

will be expected to contain roughly independent samples of $\mathcal{X}_d^2$. The chi-square Q-Q plot is a scatter plot of the empirical quantiles of the standardized Mahalanobis distances and theoretical quantiles of the chi-square distribution. The quantiles and corresponding cumulative probability values of the empirical data distribution are derived from the ordered set of sample distances

$$\left\{\mathcal{M}_D^2(\boldsymbol{y}_i;\boldsymbol{\mu}_i, K_i)_{(1)} \leq \mathcal{M}_D^2(\boldsymbol{y}_j;\boldsymbol{\mu}, K_j)_{(2)} \leq \ldots \leq \mathcal{M}_D^2(\boldsymbol{y}_k;\boldsymbol{\mu}_k, K_k)_{(m_*)}\right\}.$$

The cumulative probability of the empirical quantile is subsequently mapped to the corresponding theoretical quantile of $\mathcal{X}_d^2$. If the empirical data distribution is representative of the theoretical distribution, within effects of limited sample availability at all test locations and theoretical simplifications of the real data noise, the Q-Q pairs will roughly plot along the identity line. Deviations of the Q-Q scatter trend from the identity line can thus be used for cross-validation of the posterior predictive uncertainty estimates.

3. Application to Central Valley (CV)

Our study area is Central Valley covering approximately 20,000 square miles in central California (Figure 1). Groundwater pumping is a primary source of water support for its vast agricultural system (Bertoldi et al., 1991; Faunt, 2009; Williamson et al., 1989). For convenience of discussion, the study area is typically divided into the

northern Sacramento valley (SV) and southern San Joaquin valley (SJV). The CV groundwater basin has been delineated into several subbasins based on factors such as hydrogeologic barriers or institutional boundaries (California Department of Water Resources, 2003), which have been overlain on the CV map shown in Figure 1. The sediments of the underlying aquifer system were derived from the surrounding Sierra Nevada and the Coast Ranges. Defining stratigraphic units in the CV aquifer system has generally been difficult due to absence of distinct lithologic changes (Faunt, 2009). The SV sediments have been determined to constitute of coarse-grained alluvial sediments interbedded with localized fine-grained sediments attributed to low-energy drainage basins. In the SJV, the hydrogeologic makeup is described in terms of an upper semi-confined and lower confined aquifer zone, separated by a confining unit. Three intermixing hydrogeologic units, namely Coast Ranges alluvium, Sierran alluvial deposits, and flood-basin deposits, form the constituents of the upper semi-confined aquifer zone (Laudon & Belitz, 1991). Fine-grained alluvium, predominantly derived from the Coast Ranges, are present in the form of spatially discontinuous lenticular shapes, comprising approximately half of the volumetric fill. In contrast to the SV, within the SJV there is a distinct and spatially continuous confining unit dividing the upper and lower aquifers consisting of low-permeability clay deposits known as the Corcoran Clay. The spatial extent of the Corcoran clay is well-mapped (Figure 1) as described in Section 3.1.

While the general hydrogeologic characteristics of the CV has been well-studied, physically modeling groundwater flow in the CV encounters large uncertainty resulting from the regional and local stratigraphic variations of the hydraulic and storage properties required in flow modeling studies. We undertake a machine learning approach to address this spatial uncertainty. Specifically, our approach involves the Gaussian process methodology formulated in Section 2 to model groundwater level long-term and seasonal variability trends using hydrogeologic features.

3.1. Available Data

We describe below the well and lithological texture data sets available to us in the study area. Note that while the proposed methodology will work on irregular, non-discretized data, we perform discretization of the well data for ease of analysis with the lithologic texture data which is available in gridded manner. We choose a 1 square mile spatial resolution for every cell in the modeling grid over CV to correspond with that of the sediment texture data.

1. *Water level data*: We use water level time series data across the CV as compiled and processed by Kim et al. (2021) using groundwater well data sets obtained from the California Department of Water Resources (DWR) and United States Geological Survey (USGS). The data set consists of measurements from approximately 4,500 wells across our study area. Well screen depth information is not available. We discarded approximately half of the wells for having too few samples (less than 8) along time axis. A few wells were ignored as they clearly contained outlier samples. The well data were spatially aggregated into the modeling grid over the CV. Many grid cells, post aggregation, contained multiple individual wells, which should not be co-mingled within the time series. Additionally, there were cases where two different time-series were presented from two different data sources for one co-located grid cell, leading to confusion which measurement was to be considered for this study. To work around this issue and to apply a uniform standard across the modeling grid of study, the most temporally robust well record within each grid cell was selected. After spatial aggregation at the modeling grid resolution, data is available at approximately 1,750 spatial locations (Figure 2). Temporally, the data was available till August 2020 and we considered a rough 5 year period starting from March 2015 for this study. The well time series data was averaged at biweekly intervals. This resulted in well data aggregated into the spatio-temporal grid having 400, 220, and 132 cells along latitude, longitude and time axes respectively. Figure 2 shows the time series data at three wells. The footprint of the dry-wet seasonal cycle on the water levels can be clearly observed in the top plot. The best fitting long-term and seasonal trend (red line) to the well data is discussed further in Section 3.2.1.
2. *Lithological texture data*: Information on subsurface hydrogeology in the form of volumetric proportion of coarse and fine grained sediments (commonly referred to as lithological texture) is available on a uniformly discretized grid across the CV, as modeled in previous work by Marcelli et al. (2022). The authors generated the texture model by 3D kriging of texture observations derived from approximately 8,500 drillers logs. To address the challenges of non-stationarity in kriging, the authors divided the study area areally and vertically into several modeling domains. Kriging was performed separately across each groundwater subbasin (see Figure 1) and aggregated vertically into 13 layers. The stratigraphy of the 3D model is largely determined by

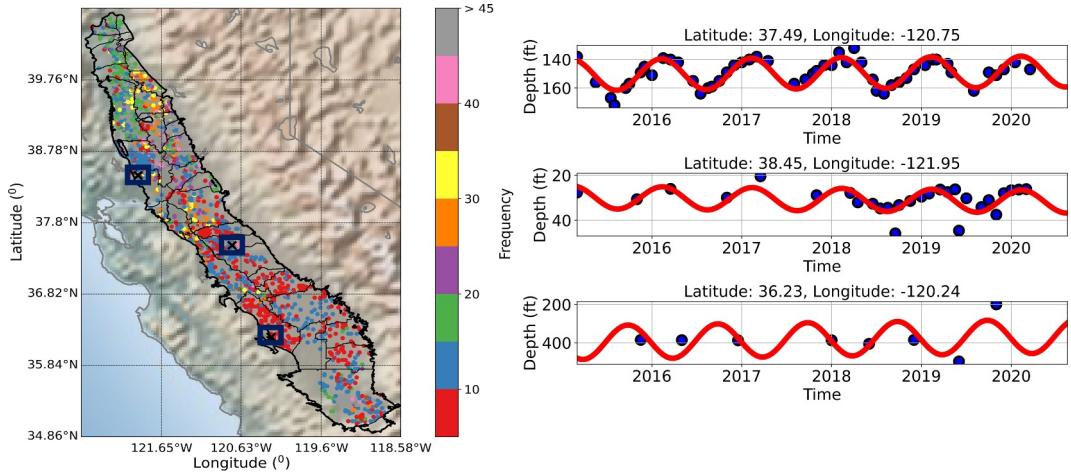

Figure 2. (left) Well locations colored by the frequency of data samples. Wells shown in the right plot are highlighted with black crosses. (right) Water levels measurements at three wells are shown in blue circles. Best fitting long term and seasonal trend line, estimated using Equation 15, is overlain in red on well data.

the structure of the Corcoran Clay, represented by layers 6–8 in the model. The remainder of the layers are divided between the upper semi-confined and lower confined aquifers.

3.2. Training Data Set Generation

The objective is to estimate the posterior predictive distribution $\boldsymbol{a}_*|\boldsymbol{y}_\tau, \hat{X}_\tau, \hat{\boldsymbol{x}}_*$ as specified in Equation 6 for all $\boldsymbol{x}_*$, in the CV grid. We randomly distributed available wells into training, validation and test sets of sizes 1,550, 100, and 100 wells respectively. We also consider a robust test set of 90 wells, which is created by removing 10 outlier wells from the original test set using the outlier detection scheme described in Section 3.3. For each evaluation set, we created target variables and input features as discussed below.

3.2.1. Prediction Variables

We seek to predict quantitative metrics of groundwater level fluctuation during the 2015–2020 study period. As discussed in Section 1, it is useful to model the seasonal and long-term water level fluctuation trends over time. In this paper, we assume that measured CV water level depths vary in 2D only as a function of spatial latitude and longitude coordinates $\boldsymbol{x}$ and time t. Additionally, we assume that water level time series data from 2015 to 2020 at each $\boldsymbol{x}$ may be decomposed as a linear model for the long-term signal and a sinusoidal model for the seasonal signal. Mathematically, water level at $\boldsymbol{x}$ varies through t as

$$u(\boldsymbol{x}, t) = a_1(\boldsymbol{x}) + a_2(\boldsymbol{x})t + a_3(\boldsymbol{x})\sin\left(\frac{2\pi t}{\lambda} + a_4(\boldsymbol{x})\right) \tag{15}$$

where, $a_1(\boldsymbol{x})$, $a_2(\boldsymbol{x})$, $a_3(\boldsymbol{x})$ and $a_4(\boldsymbol{x})$ denote the intercept, slope, amplitude and phase parameters respectively. Note that the simplifying assumption of 2D variability of water levels is made to facilitate initial evaluation of the efficacy of GP-DNN methodology for hydrological modeling. Extending proposed methodology for rigorous 3D-modeling of groundwater levels, where vertical connectivities between all 13 aquifer model layers are accounted for, is left as future work. We also chose to make the simplifying assumption of a single long-term and seasonal model across 2015–2020 as the well data is very sparsely sampled along time at several well locations (see well data frequency in Figure 2) and complex models may overfit to the data. However, note that the proposed methods can easily be extended to other complicated temporal models, for instance the B-spline integrated with multi-period sinusoidal model considered by Riel et al. (2018).

The long-term and seasonal parameter fields are estimated independently at each well by solving a linear regression problem with the corresponding time series data (see Appendix B). Figure 2 shows the modeled water level signal along with observed data at three well locations, while Figure 3 shows the estimated parameters at all

10.1029/2024JH000322

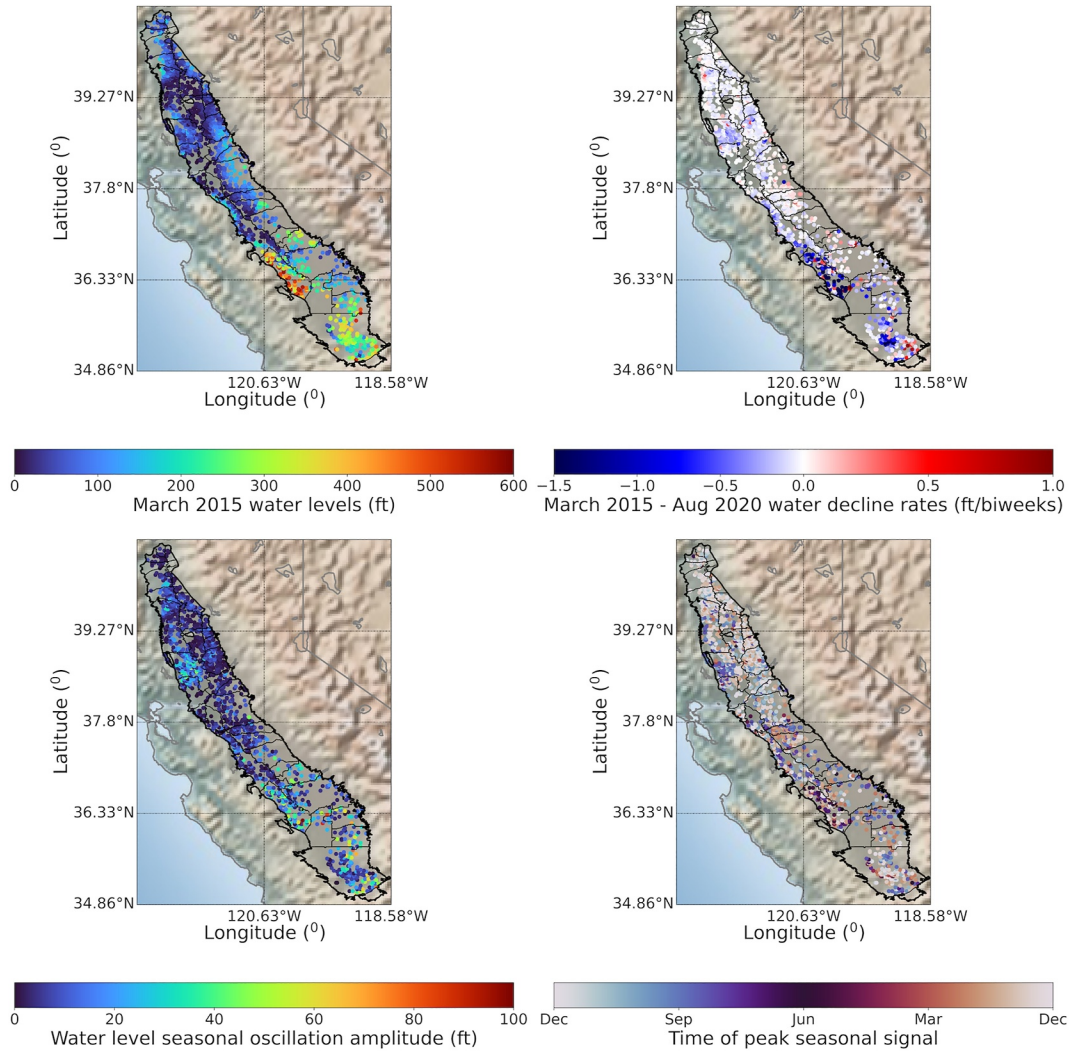

Figure 3. Water level long-term and seasonal model parameters fitted by linear regression on well time series data.

wells. The results indicate that in March 2015 (start of our analysis period), water levels were relatively deeper in the SJV as compared to the SV. During the next 5 years, wells exhibit both long-term decline and uplifts of water levels, with largest uplifts observed in the wells of the Westside subbasin and Kern county. The seasonal amplitude signal generally has high magnitudes in the southern SJV. A common feature across the long-term and seasonal trend parameters is the smoother variability in the northern two-thirds of the valley, with greater spatial heterogeneity in the southern San Joaquin basin, likely a manifestation of the underlying hydrogeologic heterogeneity as discussed in Section 3.4. Another complicating factor is data sparsity since most wells in the southern SJV contain <15 samples (Figure 2). Thus, the linear regression trend estimates are expected to be noisier. The GP methodology accounts for this noise through the estimated noise-level matrix Σ_n (see Equation 4). As discussed in Section 3.4, the hierarchical GP-DNN model correctly identifies this data uncertainty by predicting wider uncertainty intervals in the southern SJV.

It should be noted that trend data at wells cannot be expected to be Gaussian in nature, for instance, water levels depths cannot assume negative magnitudes and are skewed toward positive values. Thus, the GP methodology is not directly amenable to the water level data. A simple yet effective approach to handle this issue is to perform a normal score transform (Journel & Huijbregts, 1978), which transforms the sample data histogram of each prediction variable into the standard Gaussian distribution. In this paper, the GP regression is performed on the transformed normal variables and the regression outputs, as shown in Section 3.4, were obtained by subsequent back-transformation to replicate the original well data sample histogram.

3.2.2. Feature Variables

We consider the following features as input to the GP regression as discussed below.

1. *Geospatial coordinates* $\boldsymbol{x}$: The baseline features we consider are the latitude and longitude coordinates. The baseline GP regression employs only $\boldsymbol{x}$ as features as discussed earlier, while GP regression in extended feature space uses $\boldsymbol{x}$ and additional features as discussed below.
2. *Hydrogeological features* $\check{\boldsymbol{x}}$: Amongst the several factors that groundwater flow depends on, geologic variability of the underlying aquifer plays a crucial role. Hydraulic head evolves in 3D inside an aquifer and can be physically described through the 3D groundwater flow equation (Harbaugh, 2005),

$$\frac{\partial}{\partial x}\left(K_{xx}\frac{\partial h}{\partial x}\right)+\frac{\partial}{\partial y}\left(K_{yy}\frac{\partial h}{\partial y}\right)+\frac{\partial}{\partial z}\left(K_{zz}\frac{\partial h}{\partial z}\right)+W=S_s\frac{\partial h}{\partial t}, \tag{16}$$

where h is the hydraulic head, x, y, z denote the spatial dimensions, t denotes time, S_s is the specific storage coefficient, K_{xx}, K_{yy}, K_{zz} represent hydraulic conductivity along the spatial dimensions and W represents the flow source/sink term. For unconfined aquifers, specific yield is used as the storage coefficient. While the general groundwater flow equation presented above calculate hydraulic head in all three spatial directions, we use a 2D assumption for groundwater level variation (see Section 3.2.1). For subsurface rocks, the hydraulic and storage properties will vary depending on several factors, including the lithologies, lithological composition and microstructure (structural arrangement of the sediments and pores) of the porous rock medium (Mavko et al., 2009). In general, these properties may be measured by field well tests or laboratory tests on rock core samples. For instance, hydraulic conductivity for a core sample may be measured by a permeameter as

$$K=\frac{\Delta V}{\Delta t}\frac{Z}{Ah} \tag{17}$$

(Todd & Mays, 2005), where ΔV represents the volumetric flow of water in time Δt, Z is the thickness of the sample, A is the area of the sample and h is the hydraulic head. Similarly, specific storage coefficient is defined as the volume of water retained or released from a porous medium per unit volume of the aquifer per unit change in h. While corresponding in situ measurements are irregularly available across the CV, the 3D lithological texture model (Section 3.1) may be used as a proxy for the aquifer lithology (Faunt, 2009), hydraulic and storage properties.

In our methodology, the GP-DNN model will attempt to discover useful correlations between the observed water levels and the texture features. Given the texture model, we assume that the aquifer system rocks consists of two lithological end-members, the coarse-grained and fine-grained lithologies. For each layer of the texture model, the following features related to the depth and thickness of the lithologies are extracted. Thickness features were specifically chosen since effective hydraulic or storage properties across a sediment column will depend on the volumetric proportions of lithological end-members (Equation 17).

(a) *Coarse-grained sediment thickness*: For the i^{th} layer, the thickness of coarse-grained sediments at location $\boldsymbol{x}$ is computed as

$$z_{coarse,i}(\boldsymbol{x})=f_{coarse,i}(\boldsymbol{x})z_i(\boldsymbol{x}),\ \forall\, i,$$

where $f_{coarse,i}(\boldsymbol{x})$ denotes the volumetric fraction of coarse grained sediments at location $\boldsymbol{x}$ and layer i and $z_i(\boldsymbol{x})$ denotes the depth thickness of layer i. Directly using $f_{coarse,i}$ as a feature may be misleading to the machine learning model since the thickness of the layers may vary significantly across $\boldsymbol{x}$. We expect $z_{coarse,i}(\boldsymbol{x})$ to be informative on effective volume of the coarse-grained lithology for the layer at any given $\boldsymbol{x}$.

(b) *Fine-grained sediment thickness*: Assuming only two end-member lithologies, we compute a similar effective volume feature for the fine-grained lithologies

$$z_{fine,i}(\boldsymbol{x})=\left(1-f_{coarse,i}(\boldsymbol{x})\right)z_i(\boldsymbol{x}),\ \forall\, i.$$

(c) *Depths to layer tops*: We also incorporate the depth to each layer top as a feature. Given that the properties of the layers below the water levels will vary due to fluid presence in the pores, it is desirable that the neural network is able to extract any potential correlations between the layer top depths and the water levels. Note that the depths to layer tops of the texture model were available as measured from the mean sea level. To make the datum equivalent to the water level depths which are measured from the surface and contain effects of surface topography, surface elevations as obtained from the NASA Digital Elevation Model (NASA JPL, 2020) were added to the layer depths at each grid location.

Figure 4 shows the layer top depths, coarse and fine grained sediment thicknesses for three texture model layers. Note that we use a total of 39 hydrogeological features (3 for each of the 13 layers) in our analyses.

3.3. Training, Hyper-Parameter Tuning and Cross-Validation With Blind Wells

The baseline GP model and GP-DNN models are regressed using a normalized training set containing pairs of feature and prediction variables. The input features variables in the training set were normalized to have zero mean and unit variance per feature category, that is, latitude, longitude, coarse sediment thickness, fine sediment thickness and depths to layer tops. We underscore that for the hydrogeological features a single normalization is applied across all the model layers per feature category to preserve inter-layer feature correlations. The prediction variables were normal score transformed as discussed in Section 3.2.1, hence also have zero mean and unit variance in the training set. We briefly summarize the model training and hyper-parameter tuning with detailed description of the results presented in Appendix C. For the baseline case, the posterior predictive distribution can be derived analytically and requires no explicit training. The GP-DNN model contains a DNN in the bottom layer of the hierarchy. The DNN architecture constitutes several hidden neural network layers, each of which consists of a number of neurons with trainable weight and bias parameters $\boldsymbol{\theta}$, and a multivariate output layer. By hyper-parameter tuning as described below, the optimal DNN architecture was found to consist of 2 hidden layers with 33 neurons in each layer. The dimension of the output latent space p was tuned to be 12. A standard GP model is finally regressed in the space of DNN outputs $\tilde{\boldsymbol{x}}$. Training of the DNN parameters is performed end-to-end by stochastic gradient descent (Algorithm 1).

The GP model requires specification of hyper-parameters related to the anisotropic length-scales of input variables (see section Appendix A), observational noise levels of the target variables and the amplitudes of the kernel in K_{amp}. The GP-DNN model additionally requires hyper-parameters related to the DNN architecture, such as the number of hidden layers, and to the training algorithm. Hyper-parameter tuning is performed by finding the best fit under cross-validation. Three thousand sample sets of hyper-parameters are generated by random sampling over their range of variability. The negative log-likelihood cross-validation statistic (see Section 2.2.4) is evaluated on the validation set using the 3,000 hyper-parameter sets, and the one that optimizes the statistic is chosen. Table 1 compares the final log-likelihood statistic for the two models considered across different evaluation sets. Also shown is the root mean square error (RMSE) between the predicted posterior mean and prediction variables $\boldsymbol{y}$. The training set is the set of wells used for training of the GP-DNN model parameters, while the validation set is the set of wells used for hyper-parameter tuning. The test set is kept blind completely and used only for final validation. We also consider a robust test set by removing 10 outlier wells from the test set as described below. Note that the validation statistics are computed on the normalized evaluation sets. Given that the DNN has a 1D architecture and the GP posterior predictive distributions are computed analytically, end-to-end training of the DNN model for each hyper-parameter scenario takes about a minute to complete on a machine having 32 GB random-access memory (RAM) with a single 32 GB graphics processing unit (GPU). For effectively tuning larger sized deep learning models, efficient hyper-parameter tuning frameworks based on Bayesian optimization (Akiba et al., 2019) may be considered.

While GP-DNN exhibits slightly higher RMSE in the mean estimates, the GP-DNN regression outperforms the conventional GP regression in terms of the uncertainty estimates as evidenced by the higher likelihood of the well data under the estimated posterior predictive distribution. In Table 1, we have decomposed the negative log-likelihood values according to the R.H.S. of Equation 12 to aid interpretability. As expected, the baseline GP model offers simpler models and performs better in terms of model complexity as evidenced by the relatively lower $\frac{1}{2}\log|K|$ values. However, the GP-DNN model performs significantly better in explaining the data

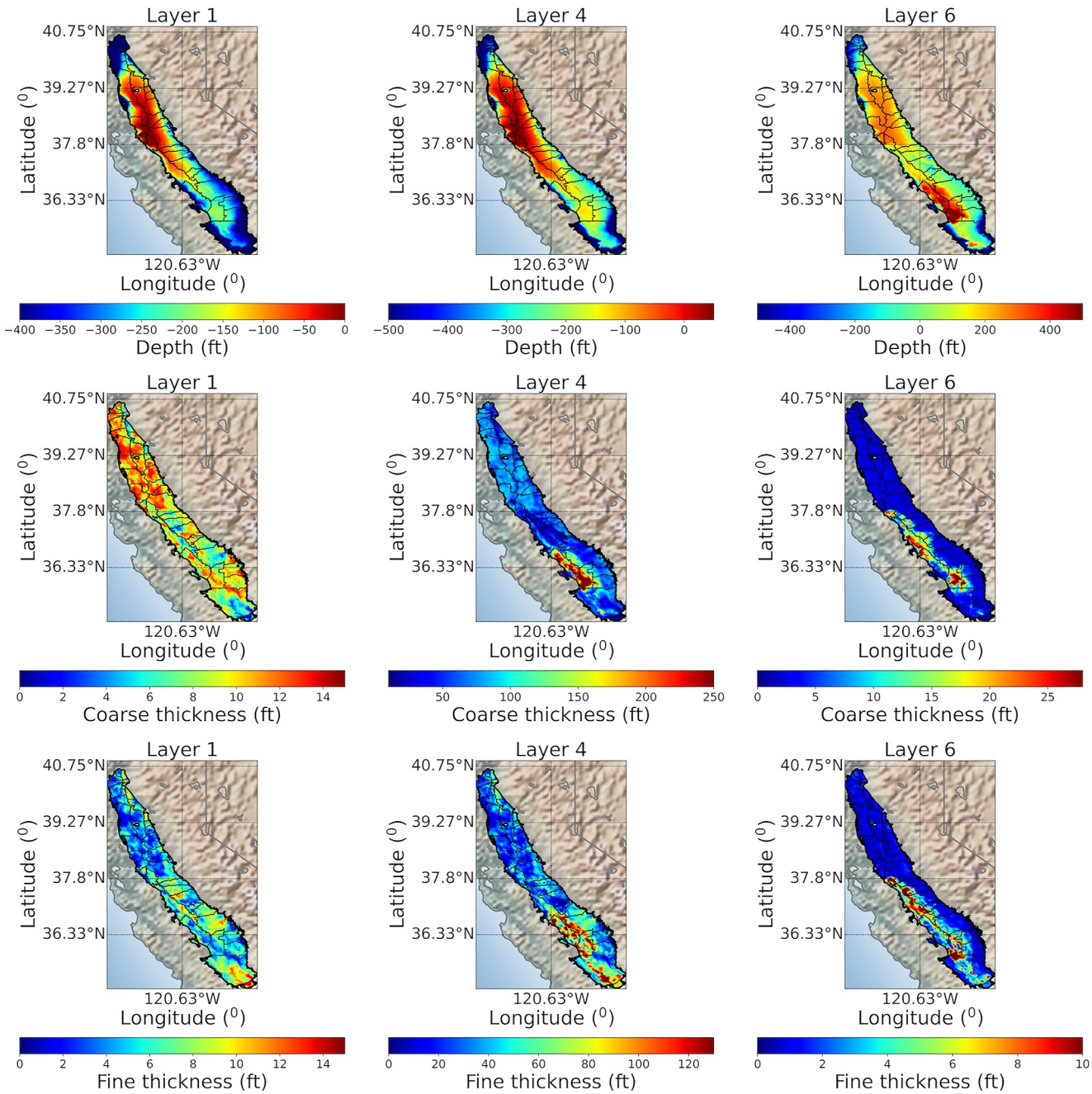

Figure 4. Depth to layer tops (top row), coarse-grained sediment thickness (middle row) and fine-grained sediment thickness (bottom row) for three different texture model layers. Top of layer 6 corresponds to top of Corcoran clay.

variability as demonstrated by the significantly lower standardized Mahalanobis distances of the evaluation set from the predicted probability density. We underscore that the objective of the paper is not to find the best mean model, rather derive an informative predictive posterior that is able to robustly quantify the uncertainty due to the real data irregularity and noise. We use chi-square Q-Q plots next to further bolster the claim that this is achieved with the GP-DNN posterior estimates.

In Figure 5, we compare chi-square Q-Q plots to determine the fidelity of the GP and GP-DNN uncertainty estimates toward explaining the uncertainty exhibited in the blind well test set $\mathcal{T} = \{\boldsymbol{x}_{*1}, \boldsymbol{x}_{*2}, \ldots, \boldsymbol{x}_{*m_*}\}$. Consider

Table 1
Training and Cross-Validation Statistics (Defined in Section 2.2.4) Computed on Different Normalized Evaluation Sets

	Baseline GP		GP-DNN	
Evaluation set	Posterior predictive − log-likelihood ($\frac{1}{2}\mathcal{M}_D^2 + \frac{1}{2}\log\|K\| + \frac{dm_*}{2}\log 2\pi$)	RMSE	Posterior predictive − log-likelihood ($\frac{1}{2}\mathcal{M}_D^2 + \frac{1}{2}\log\|K\| + \frac{dm_*}{2}\log 2\pi$)	RMSE
Training set (1,550 wells)	5211292.05 (5222822.08–17205.39 + 5675.36)	0.64	12217.34 (12009.6– 5467.62 + 5675.36)	0.45
Validation set (100 wells)	2784.99 (3017.2–599.79 + 367.58)	0.77	446.51 (256.65–177.72 + 367.58)	0.85
Test set (100 wells)	3058.42 (3302.32–611.48 + 367.58)	0.75	439.43 (259.16–187.31 + 367.58)	0.82
Robust test set (90 wells)	2024.80 (2245.71–551.73 + 330.82)	0.64	345.16 (189.99–175.65 + 330.82)	0.68

the hypothetical scenario in which observations $\boldsymbol{y}(\boldsymbol{x}_{*i}) \in \mathbb{R}^{d=4},\ \forall \boldsymbol{x}_{*i} \in \mathcal{T}$, are true samples from the predictive posterior distribution $a(\boldsymbol{x}_{*i})|X_\tau, \boldsymbol{y}_\tau, \boldsymbol{x}_{*i}$ (Equation 6). The Q-Q plot would then result in an identity relation as discussed in Section 2.2.4. We verify this claim in the left column of Figure 5 where the empirical quantiles on the y-axis are estimated using 100 random samples from $a(\boldsymbol{x}_{*i})|X_\tau, \boldsymbol{y}_\tau, \boldsymbol{x}_{*i},\ \forall \boldsymbol{x}_{*i} \in \mathcal{T}$. The Q-Q plot shows an almost perfect identity relation. In our problem setup, only one observation per $\boldsymbol{x}_{*i}$ is available. The effect of this limited sample size is explored in the second column from left in Figure 5 where one sample per $a(\boldsymbol{x}_{*i})|X_\tau, \boldsymbol{y}_\tau, \boldsymbol{x}_{*i}$ is utilized. The Q-Q scatter plot shows minor deviations around the identity line, with appreciable deviations generally observed for $\mathcal{M}_D^2 > 10$, with the value 10 corresponding to the 96% quantile of $\mathcal{X}_{d=4}^2$ distribution. As highlighted by R. A. Johnson and Wichern (2007), such Q-Q deviations at tail ends of the distributions become exacerbated due to limited sample size.

In the second column from right of Figure 5, the empirical quantiles on y-axis are derived using the real observations $\{\boldsymbol{y}(\boldsymbol{x}_{*i});\ \forall \boldsymbol{x}_{*i} \in \mathcal{T}\}$ available at the blind test set wells. The GP-DNN Q-Q scatter points roughly plot along the identity line for $\mathcal{M}_D^2 < 8$, with the value 8 corresponding to the 90% quantile of $\mathcal{X}_{d=4}^2$ distribution. Few clear outliers with respect to the predictive posterior are also observed. The deviation from the identity relation is severe for the baseline GP Q-Q plot. Note that many empirical samples of $\mathcal{M}_D^2$ (33 out of 100 wells) even fall above the 99.99% quantile of $\mathcal{X}_{d=4}^2$ distribution which corresponds a value of 23.5. This highlights that it is extremely unlikely that these samples belong to the baseline GP predictive posterior. Two potential factors that could be contributing to the lack of a clear one-to-one correspondence for the baseline GP model are (a) inaccurate regression estimates for the posterior uncertainty, and (b) real data noise. With regards to the latter, certain additional wells with suspect data were identified from the Q-Q plot outliers. For instance, the well with $\mathcal{M}_D^2 \approx 40$ (see the bottom plot on second column from right) has about 11 data samples indicating water levels declined by 170 feet from 2015 to 2017, and regained back 170 feet from 2017 to 2019, leading to a fitted seasonal oscillation amplitude value of 81 feet. This seems a clear data outlier considering the general distribution for amplitudes observed in Figure 3. To minimize impacts of similar outlying data samples for the Q-Q plot analysis, we created an additional test set, termed the robust test set $\mathcal{T}_R$ as described below.

We conducted an additional outlier detection exercise by utilizing the robust sample covariance based Mahalanobis-distance outlier detection as proposed by Rousseeuw and Van Driessen (1999). We underscore that this outlier detection is solely based on the observed data samples and does not include any regression model covariance estimates. Specifically, the sample mean and robust sample covariances are determined from the observations of $\boldsymbol{y}$ available at all training, validation and test set wells. Subsequently, the Mahalanobis distance of samples of $\boldsymbol{y}$ are computed using the sample mean and robust sample covariance matrix. Any test well observation $\boldsymbol{y}$ which exceeds the outlier detection threshold, set to be the 99% quantile of $\mathcal{X}_{d=4}^2$, are flagged as outliers with respect to the sample data distribution. Following this procedure, 10 test set wells were discarded for having data outliers. In the rightmost column of Figure 5, we compare the Q-Q plots of baseline GP and GP-DNN models obtained using $\mathcal{T}_R$. While the linearity of the Q-Q trend slightly improved (compare R^2 coefficient of determination for the fit of the data to the blue line), the baseline GP model still exhibits significant deviations from the identity relation reinforcing the claim that baseline GP model yielded inaccurate posterior uncertainty estimates.

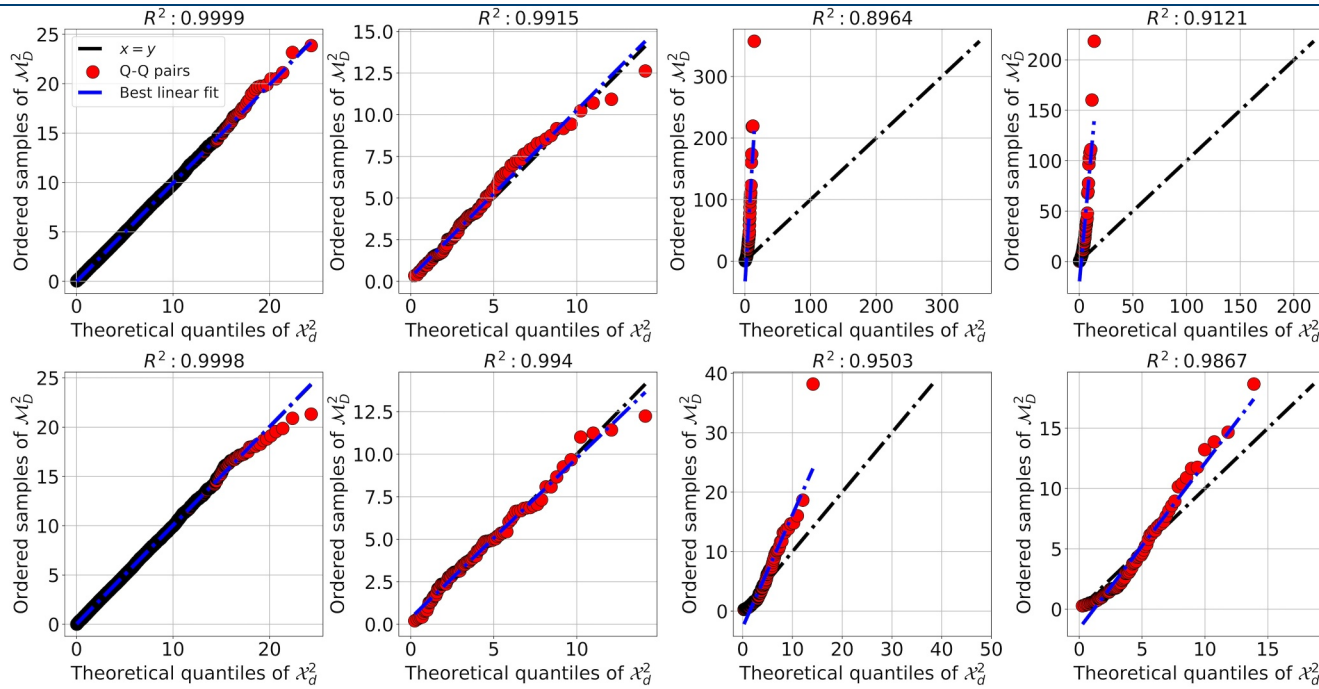

Figure 5. Chi-square Q-Q plots for testing GP (top row) versus GP-DNN (bottom row) predictive posteriors. Empirical quantiles in the left and second from left columns are derived using 100 and 1 random samples respectively from predictive distribution at each well location in the test set. Empirical quantiles in the second from right and right columns are derived using real data in the test and robust test sets respectively. The R^2 coefficient of determination of the best linear fit (blue line) is listed on top.

For the GP-DNN model, the R^2 coefficient of the best linear fit to the Q-Q trend is 98.67% indicating that the shape of the empirical and theoretical distributions are practically identical. Approximate one-to-one correspondence between empirical and theoretical distributions are observed up to the 90% quantile (≈ 8 for $\chi^2_{d=4}$). Noticeable mismatches between the empirical and theoretical quantities especially occur beyond the 90% quantile of $\chi^2_{d=4}$. Such deviations are expected given the limited sample size at each test location as well as due to the simplified modeling assumptions made regarding the data noise, for example, spatially uniform noise variance.

3.4. Results

In this section, we will discuss estimated water level trends and associated uncertainties in the CV during 2015–2020, as inferred by two GP regression models. We show that the mean field estimates from both models show somewhat comparable spatial variability. However, the results interpretation will be primarily focused on the GP-DNN results since corresponding posterior predictive distribution is able to explain the held out test set with higher log-likelihood statistics and robust predictive posterior uncertainty estimates as discussed in the previous section. The GP-DNN model is able to model the uncertainty more reliably because of its ability to handle non-stationary data in the latent space as shown in Section 3.4.3.

3.4.1. Mean Groundwater Levels During 2015–2020

Mean water level trends predicted by the GP-DNN model are shown in Figure 6. In March 2015, water levels were predicted to be significantly shallower (upto 50 feet from surface) in the SV. The SJV, on the other hands shows greater variability. Along the north-western flank of the SJV (East Conta Costa, Tracy, Delta Mendota, and western edges of Eastern San Joaquin, Modesto, Turlock and Merced subbasins), water levels within 50 feet from surface are observed. Water levels deeper than 100 feet from surface are predicted along the eastern boundary and southern half of the SJV. The deepest water levels (>200 feet) are observed across several isolated regions in the Modesto, Turlock, Chowchilla, Madera, Westside, Kaweah, Tule and Kern counties. The spatial continuity of the patterns manifested in well data (Figure 3) has been preserved in the estimated mean fields. Larger-scale spatially correlated structures are observed in the central SV and northern SJV, while rapidly varying spatial patterns (short length-scales) are observed in southern SJV area. In a previous study by Gualandi and Liu (2021), shallow aquifer

10.1029/2024JH000322

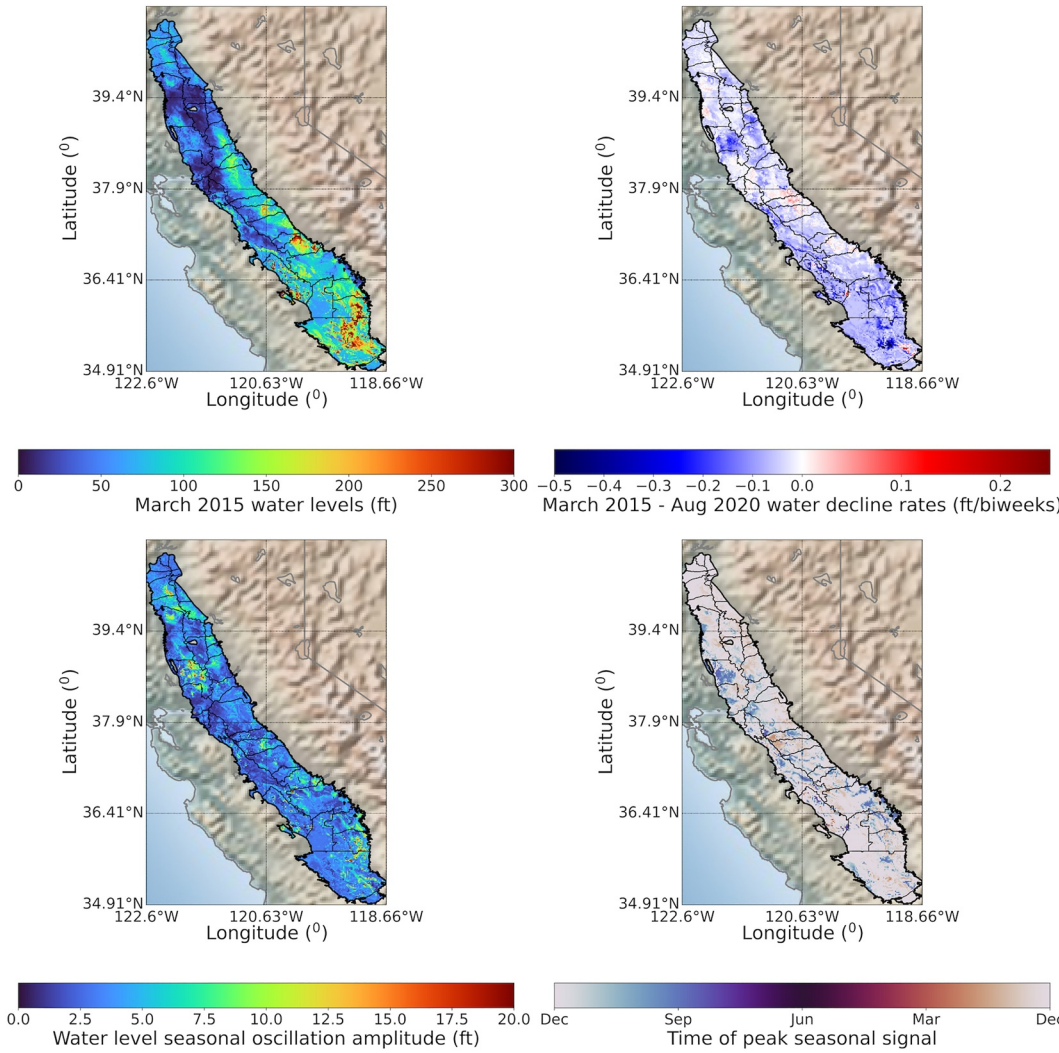

Figure 6. Mean of the posterior predictive distribution of water level long-term and seasonal trend parameters predicted by GP-DNN regression. The plots, clockwise from top left, correspond to $a_1(\boldsymbol{x})$, $a_2(\boldsymbol{x})$, $a_4(\boldsymbol{x})$ and $a_3(\boldsymbol{x})$. For $a_2(\boldsymbol{x})$, blue indicates rising water levels.

(aquifer layers above the clay confining unit) processes in the SJV were found to be contributing factors to these short length-scale variations in aquifer responses.

During 2015–2020, water levels have exhibited both positive and negative decline trends, with a large proportion of the locations showing sustainable changes in groundwater levels. On average, 98% of locations in the CV grid have had moderate fluctuations in the groundwater levels, ranging between ±12 feet during 2015–2020 (decline rates ranging between ±0.1 ft/biweeks). 11% of the locations have witnessed uplifts exceeding 12 feet, while 1% of the locations underwent declines exceeding 12 feet. While CV groundwater reservoirs continually experienced groundwater loss during the 2012–2015 drought (P. W. Liu et al., 2022; Ojha et al., 2019), our results indicate that, on average, there were few locations with large declines in groundwater (>25 feet) during 2015–2020. This is likely due to the exceptionally wet years of 2017 and 2019 that have resulted in partial, localized recovery of water levels as reported in several studies (California Department of Water Resources, 2017, 2019). However, note that these short recharge periods have been interspersed with prolonged periods of drought, in what has been termed as a megadrought, and current CV groundwater levels in general lie significantly below the pre-2006 drought levels (P. W. Liu et al., 2022). Approximately 63% of the 11% CV locations that witnessed appreciable uplifts in water levels include locations in the Tulare Lake hydrological basin (TLHB; includes Westside, Kings, Kaweah, Tulare Lake, Tule and Kern county subbasins). These results align with the California

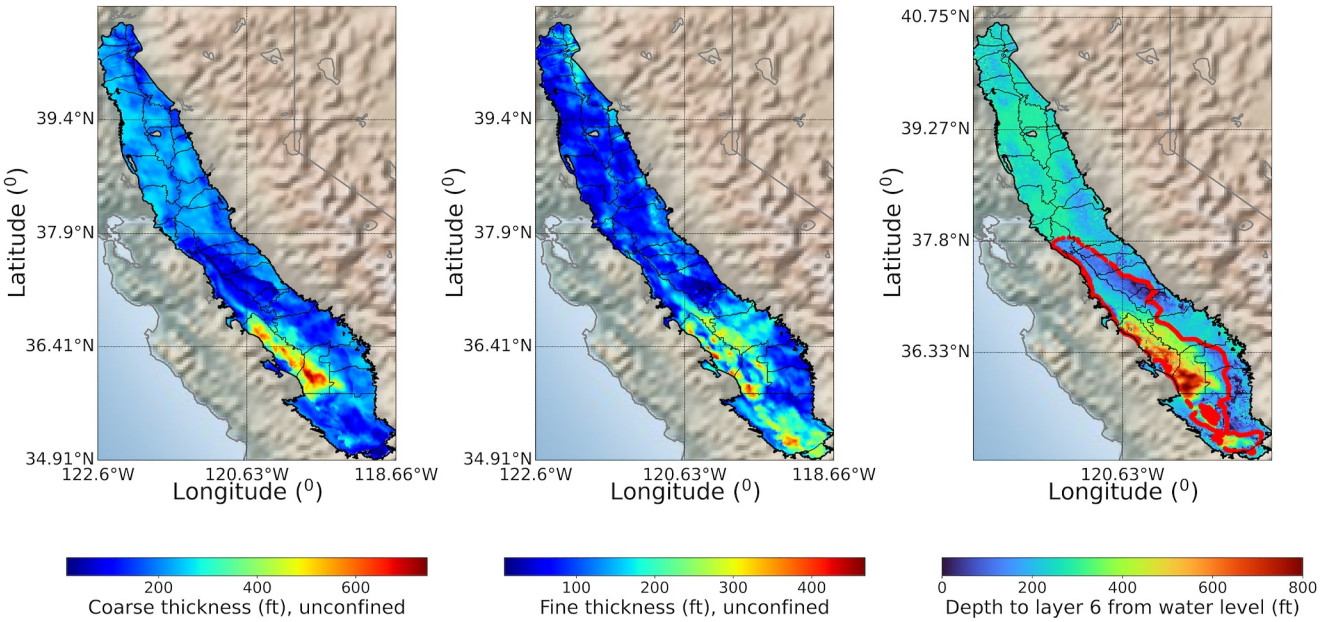

Figure 7. Total thickness of coarse (left) and fine grained sediments (middle) in the upper semi-confined aquifer (texture model layers 1–5). (right) Depth thickness between predicted March 2015 water level mean and the top the 6^{th} layer. Extent of Corcoran clay is shown in red.

Department of Water Resources (2019) report that shows several wells in the TLHB observed uplifts during Spring 2016–Spring 2019, with 31% of the 624 wells logged ranging between 5 and 25 feet uplifts and 25% exceeding 25 feet. Neely et al. (2021) also presented similar observations while studying groundwater depletion related surface deformation with remote sensing data in CV. They observed strong surface uplift in Westside during the 2017 wet year potentially due to the above average aquifer recharge in that year. We hypothesize two possible reasons for these observed uplifts in water level mean fields.

1. *Underlying hydrogeology*: Using the sediment thickness features presented in Section 3.2.2, we observed that the western flank of the TLHB contains some of the thickest coarse-grained sediment columns in the semi-confined aquifer zone. Shown in Figure 7 are the total thickness of coarse and fine-grained sediments in the upper semi-confined aquifer (layers 1–5 of the CV texture model). It may be observed that the thickness of the semi-confined coarse grained sediments in western TLHB ranges within ≈400–700 feet, significantly larger than any other region of the CV. Thick fine-grained sediments were also estimated to exist in western and southern TLHB. Also plotted is the depth to top of layer six of the texture model, corresponding to the top of the Corcoran Clay where it exists, from the GP-DNN predicted mean water level. Comparing with Figure 6, mean water levels in western TLHB are predicted to stay mostly between 50 and 400 feet below the surface and 200–800 feet above the Corcoran Clay. Note that in comparison the mean water levels in the SV and northern SJV exist at shallower depths (≈0–100 feet). Given that (a) the western TLHB semi-confined aquifer is a structural trough with thick coarse grained sediments and (b) coarse-grained sediments have higher storage and hydraulic conductivity, it is possible that the higher influx of water during wet years 2017 and 2019 resulted in preferential recharge of the western TLHB aquifers. This preferential recharge effect in western TLHB has also been confirmed in independent component analyses of InSAR time series data conducted by Gualandi and Liu (2021). As mentioned beforehand, modeling limitation of handling groundwater flow in 2D would also likely introduce dimensionality issues, such as not considering vertical soil infiltration. Further, other factors such as varying groundwater pumping rates, delayed pressure dissipation across the Corcoran clay, structural barriers or pathways inside the stratigraphies considered in the CV texture model and the spatial heterogeneity were also not considered in this study. Many of these variables are unknown or known with large uncertainty in the CV, impeding building a robust physical 3D model for groundwater flow as discussed earlier.
2. *Data sparsity and noise*: Well data from the TLHB were particularly sparse as compared to the rest of the study area (Figure 2), with nearly all of the wells having <15 samples during 2015–2020. Given that we assumed a

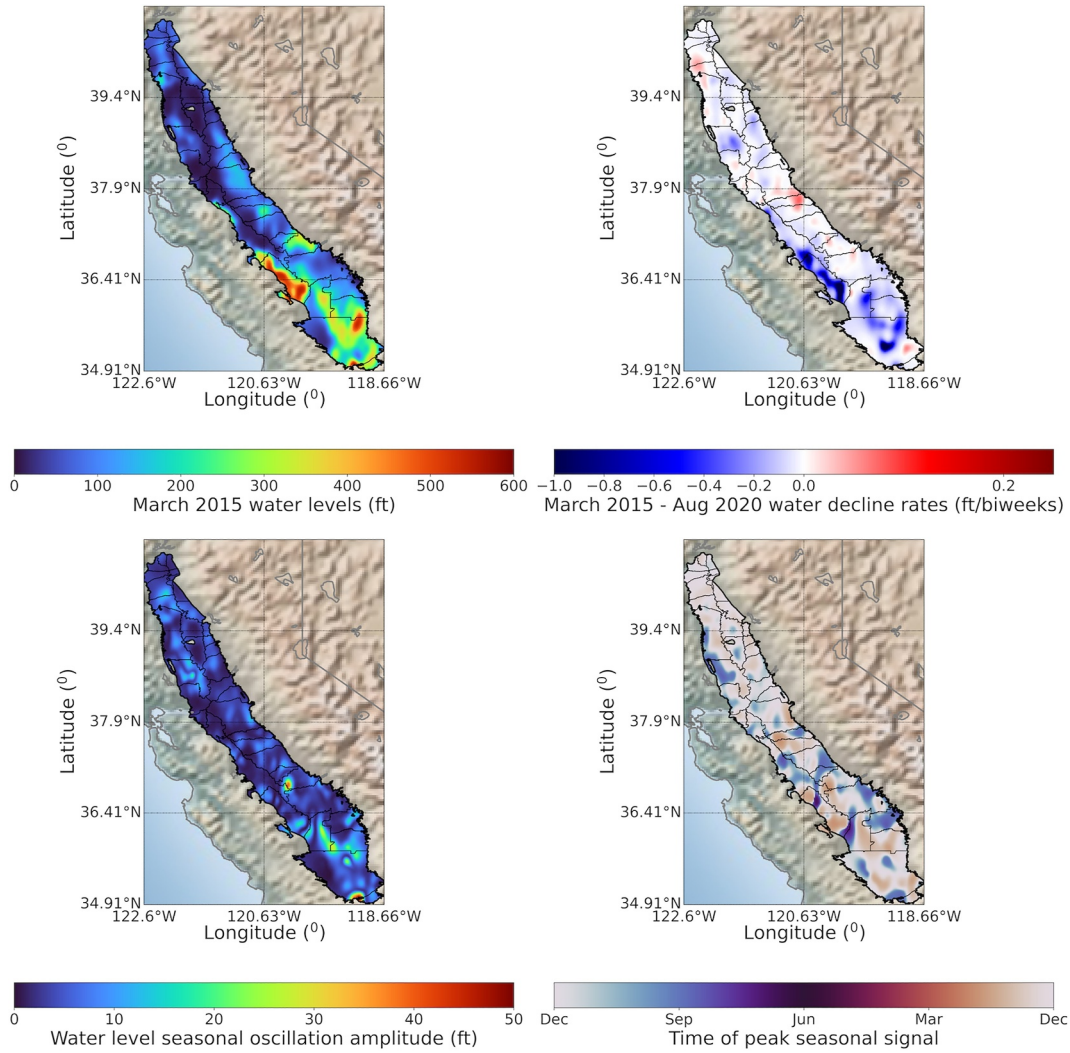

Figure 8. Mean of the posterior predictive distribution of water level long-term and seasonal trend parameters predicted by baseline GP regression.

single linear model for the long-term signal over the 5 year period, data aliasing effects could have potentially led to noisy estimates of the water level trends at wells in the TLHB.

Given the large uncertainty associated with the hydrogeology and well-data, it is desirable to quantify prediction uncertainty. In the next section, we discuss how the GP approach allows deriving the full posterior predictive distribution and the ability to simulate equiprobable estimates of the water level trends.

Regarding the mean seasonal signal, 95% of the locations are predicted to have small (<10 feet) seasonal peak-to-peak oscillations. In the SV, most of the larger oscillations are observed in Red Bluff, Corning, Vina, North Yuba, South Yuba and Yolo subbasins. In the SJV, larger amplitude seasonal variability occur regionally in the Modesto, Turlock, Chowchilla, Madera, Tule and Kern counties. Based on the phase delay field, timing of the peak seasonal oscillations are predicted to occur between October to April at 93% of the CV locations. The peak seasonal signal is controlled by various factors such as groundwater production, precipitation, snow runoff and ease of surface water infiltration into the aquifer (Riel et al., 2018). The predicted mean October to April peak generally correlates with wet months of the seasonal cycle and thus decreased groundwater production in the valley. Note that accurately estimating the timing of the peak seasonal signal will be challenging given the well data are very sparsely sampled. The GP-DNN posterior estimates appropriately capture this modeling uncertainty as discussed in Sections 3.3 and 3.4.2.

10.1029/2024JH000322

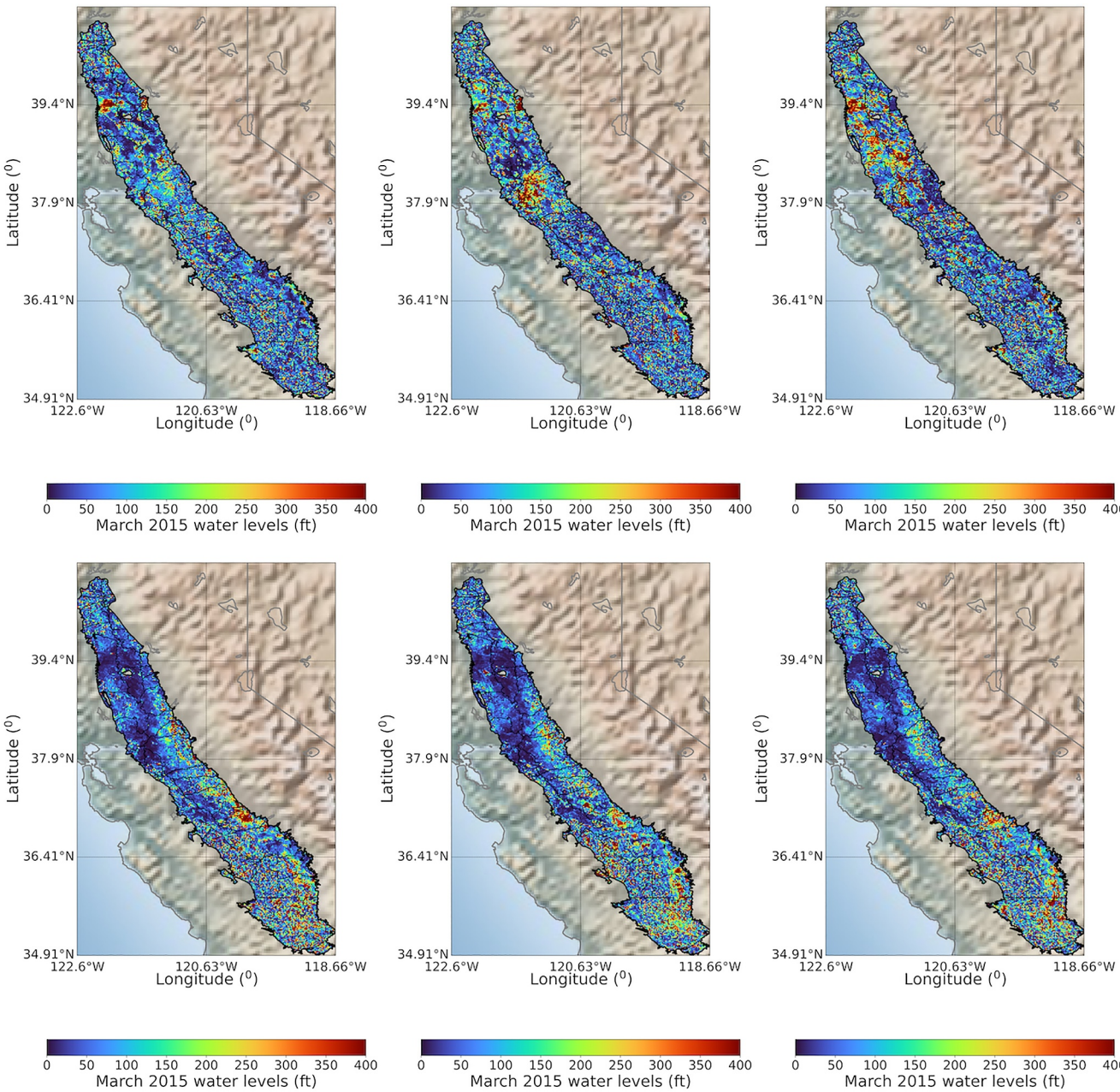

Figure 9. Three random realizations of water levels in March 2015, sampled from the GP-DNN prior (top row) and posterior (bottom row) predictive distributions.

Figure 8 shows the mean fields predicted by the baseline GP regression approach. While the fields generally show coarse correlation with the GP-DNN mean (Figure 6), the model in general tends to predict stationary correlations in the study area. This is immediately apparent in the March 2015 water levels in Westside and Tulare Lake subbasins. The rough variability manifested in the well data (Figure 3) have been smoothed given a stationary kernel is used to smooth the well data. Similar high amplitude artifacts in southern SJV are also observed in the seasonal amplitude map. The GP-DNN model, on the other hand, does not suffer from this stationary smoothing limitation and is able to capture both large-scale and fine-scale variability in the data as shown previously.

3.4.2. Uncertainty and Non-Stationarity in Groundwater Levels

Random realizations of March 2015 water levels from the GP-DNN distribution prior to well data conditioning (Equation 3) are shown in Figure 9. Non-stationary spatial heterogeneity of the simulated patterns across the

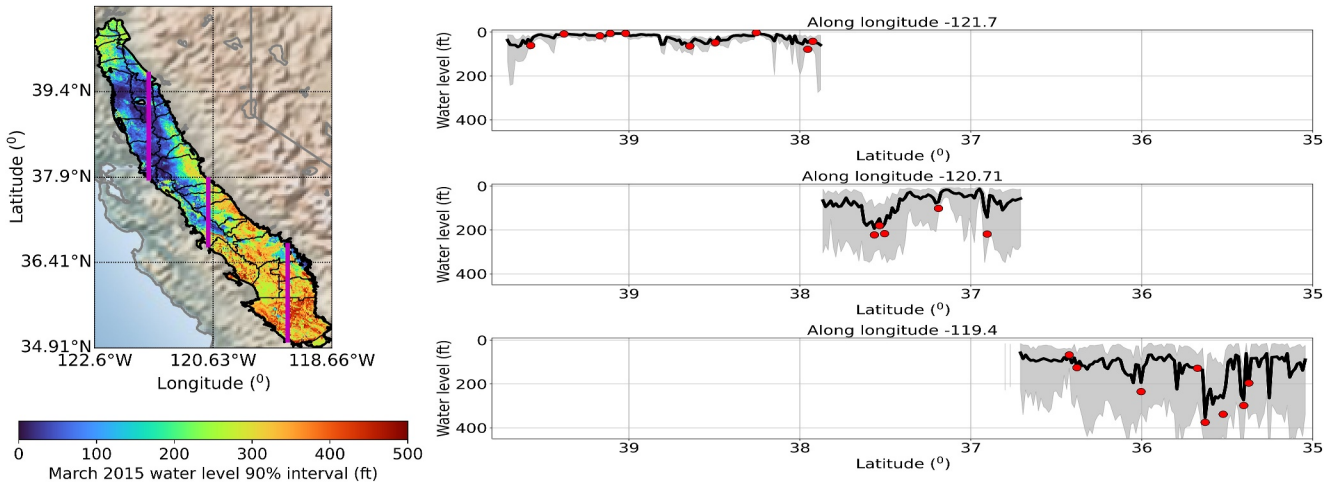

Figure 10. (left) P10-P90 uncertainty interval of March 2015 water level computed using 500 posterior predictive samples from GP-DNN model. (right) GP-DNN predictions along three different longitude transects highlighted in magenta on the left plot. Black line: posterior predictive mean; gray: P10-P90; red circles: well data.

valley is immediately apparent. Starting from central SV (Colusa, Butte and Vina subbasins) till the northern border of the SJV (Tracy and Eastern San Joaquin subbasins), large length-scale structures were simulated. Variability of water levels occurs along medium range structures in regions without the Corcoran clay confining unit such as the eastern flanks of the Modesto, Turlock, Merced, Madera, Kings, Kaweah, Tule and Kern County subbasins. On the other hand, regions located above the confining unit exhibit very fine length-scale variations, likely related to shallow semi-confined aquifer responses. The prior predictive uncertainty was conditioned to training data, yielding the posterior predictive distribution (Equation 6). Posterior predictive realizations are also shown in Figure 9. Similar to the prior realizations, the GP-DNN posterior predictive samples exhibit non-stationary spatial patterns across the modeling domain. The self-consistency of the three posterior predictive samples stands in contrast with the more highly variable samples drawn from the prior. This indicates that conditioning is, in many places within the domain, effectively constraining the prior when Equation 6 is applied.

Figure 10 shows the sample standard deviation associated with March 2015 water level mean, computed using 500 random samples from the GP-DNN posterior predictive distribution. The corresponding results for baseline GP case are shown in Figure 11 for comparison. Result interpretation is mostly focused on GP-DNN results as it was shown previously that baseline GP model predictions are inaccurate (see cross-validation in Section 3.3). We observe tight uncertainty intervals in central to southern SV and in most areas in the north-western SJV. Wider uncertainty intervals are primarily predicted for locations overlying the clay confining unit and along the northern and eastern domain boundaries. Note that the posterior predictive covariance (Equation 6) depends only on the training covariance matrix $\tilde{K}_{\tau\tau}$, testing covariance matrix $\tilde{K}_{**}$ and the training-testing covariance matrix $\tilde{K}_{\tau *}$. From Equation 7, it follows that the GP-DNN uncertainty predictions are in accordance with distances from training and

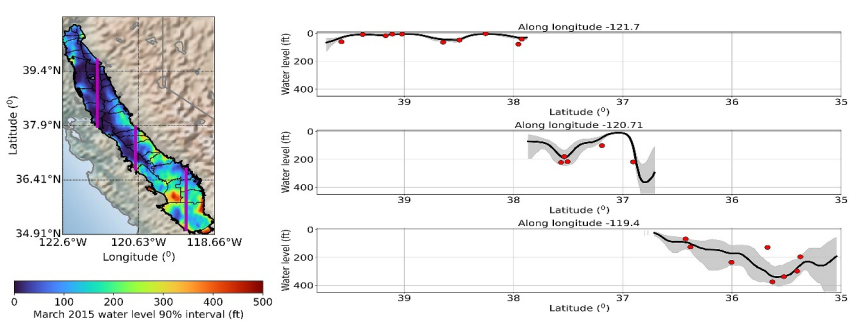

Figure 11. (left) P10-P90 uncertainty interval of March 2015 water level computed using 500 posterior predictive samples from baseline GP model. (right) Baseline GP predictions along three different longitude transects highlighted in magenta on the left plot. Black line: posterior predictive mean; gray: P10-P90; red circles: well data.

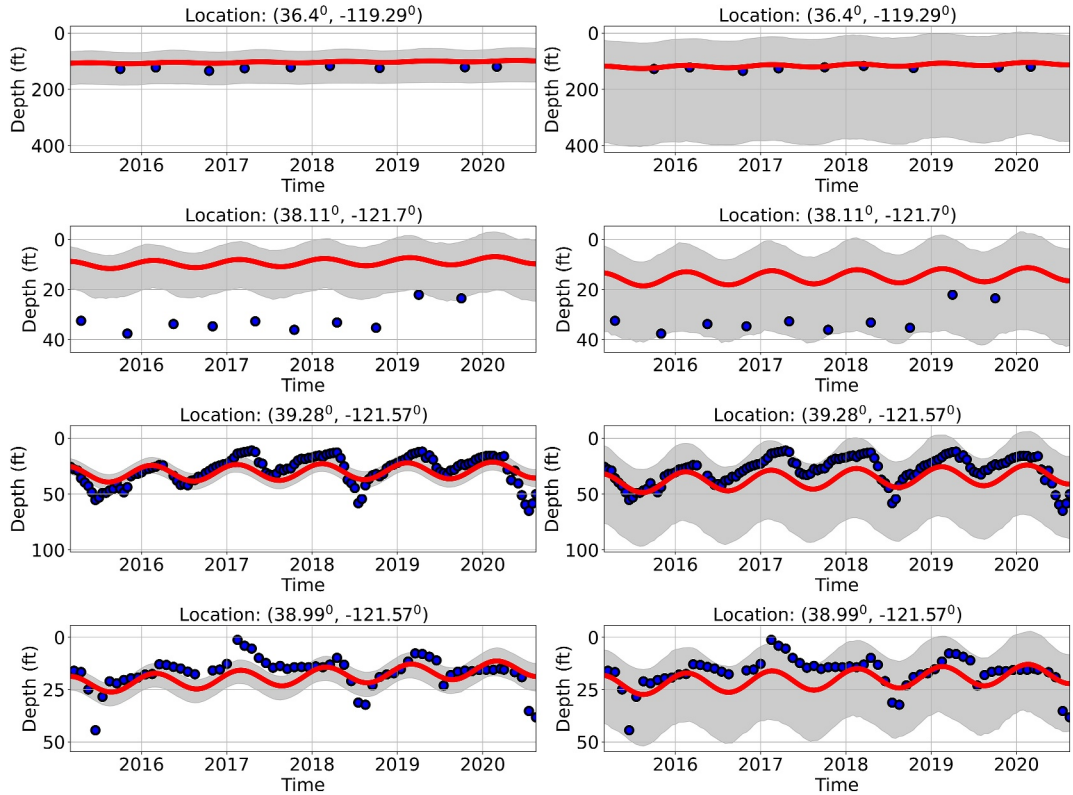

Figure 12. Modeled mean water level depths (red line), P10-P90 uncertainty intervals computed using 500 posterior predictive samples and observed data at four blind wells predicted using baseline GP (left column) and GP-DNN (right column) models.

testing locations as mapped in the latent space. Locations above the Corcoran clay have significant differences in the semi-confined aquifer structural and lithological properties, thus leading to greater separation in the latent space (see Section 3.4.3) and consequently larger uncertainty. We underscore that the well data itself shows very rapid and noisy variability in this area (which could be due to hydrogeological heterogeneity, observational noise and other factors as discussed in Section 3.4.1). As described previously, we have plotted the posterior predictive mean and 500 samples along three longitude transects and overlain the well data. Visually, well data from the northern CV show wider correlated patterns as compared to the south where the correlation falls off very rapidly with distance. This is especially true of the bottom transect passing through Westside subbasin where we see fluctuations ranging within ±300 feet roughly across a distance of 60 miles. The GP-DNN model is able to replicate this abrupt data variability within the modeling domain by latent space reconfiguration and does not force the simulations to be overly smooth, which is a well-documented limitation associated with kriging type models for modeling geological heterogeneity (Linde et al., 2015). Figure 11 shows the effect of smoothed mean estimates for the baseline GP model. Also, note that predicted uncertainty is driven by conditioning data proximity as expected. However, a limitation with the GP-DNN regression is that some of the posterior predictive samples demonstrate very noisy variability in the regions above the Corcoran clay. In Section 4, we discuss this issue in greater detail and propose ideas for future work. The linear model shown in Equation 15 may be used to obtain a spatio-temporally continuous mean water level and associated variance. Figure 12 compares the modeled mean water level and P10-P90 uncertainty intervals computed at four selected blind wells for the two regression models under study. While the mean water level predictions are more or less similar, the uncertainty intervals may vary significantly across the two wells. The GP modeling may lead to overly confident uncertainty estimates, leading to the real data lying outside the predicted P10-P90 intervals (see second well from top in Figure 12). By predicting wider uncertainty intervals in some cases, the GP-DNN model is able to appropriately capture the uncertainty existing due to data unavailability and noise.

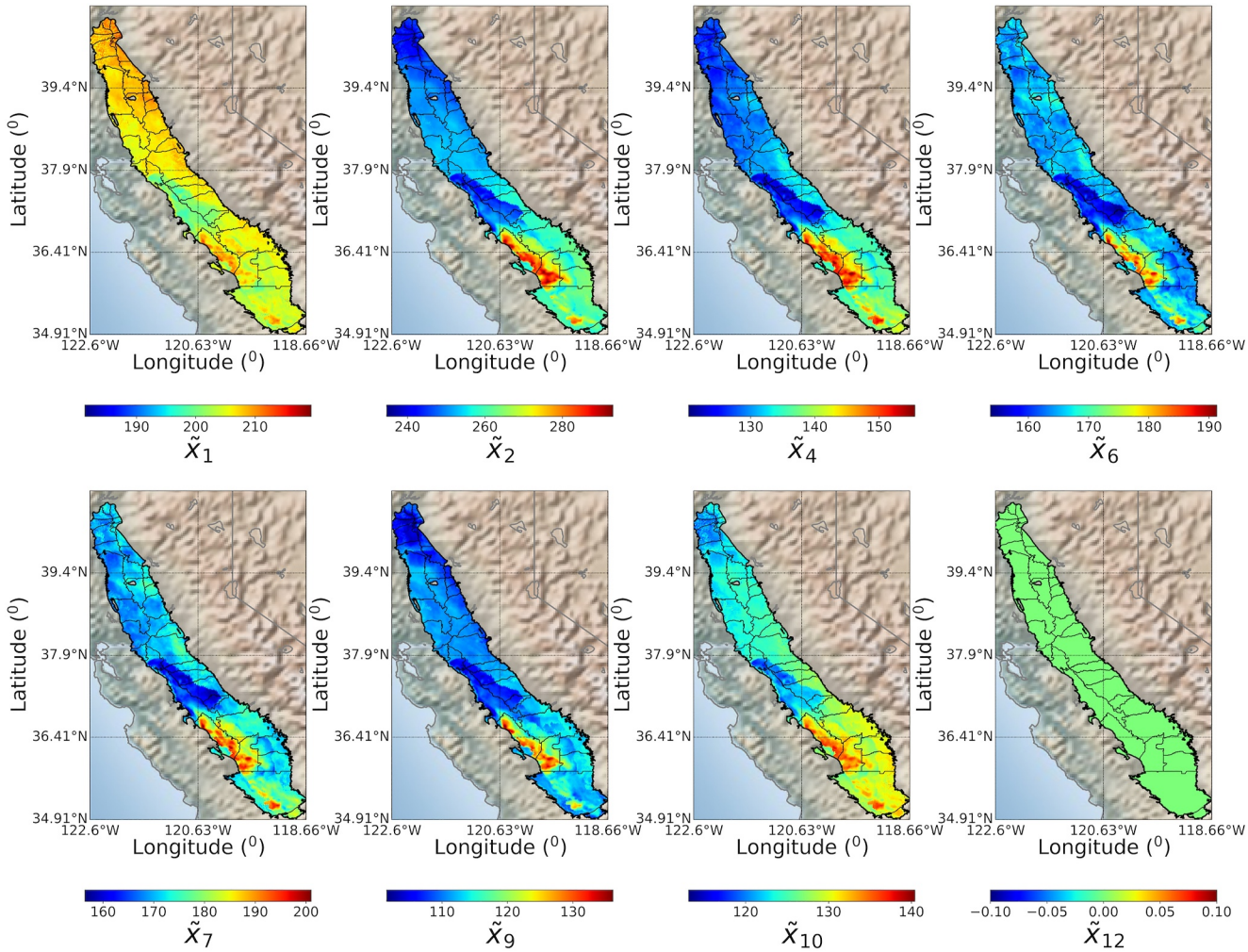

Figure 13. Selected components of the DNN output latent space $\tilde{\boldsymbol{x}}$.

3.4.3. GP-DNN Latent Space Interpretation

While deep learning models provide the flexibility to approximate very complex non-linear relations, physically explaining the model predictions, for instance understanding how the input features get propagated through the model into outputs, has remained a challenge and is an active area of research (Samek et al., 2021). In this paper, deep neural networks were used to transform the hydrogeological features into latent features. To explain the latent space features, we perform dimension reduction of the latent space and present visual analyses of the variability of aquifer and aquitard sediment thicknesses in the latent space as described in this section.

The dimensionality of the latent space, controlled by the number of output nodes of the DNN, was treated as a hyper-parameter and tuned to 12 by cross-validation. In Figure 13, we show selected components of $\tilde{\boldsymbol{x}}$ in map format, that is, plotting them versus $\boldsymbol{x}$. While it is not immediately straight-forward to quantitatively interpret the latent fields from a hydrogeological perspective, we put forth some qualitative observations. The footprint of the Corcoran clay unit can be observed along the majority of the latent dimensions. Many of the displayed latent fields show similar high amplitude patterns in the Westside, Tulare Lake and Kern County subbasins. We found that these patterns exhibit a high correlation with the coarse and fine grained sediment thicknesses in the upper semi-confined aquifer. Similar high amplitude patterns may be observed for the sediment thicknesses in the upper semi-confined aquifer (Figure 7). High amplitude structures in $\tilde{x}_6$ and $\tilde{x}_7$ (for instance in North and South Yuba subbasins) correspond with areas with thick clay columns in the lower semi-confined aquifer zone. These

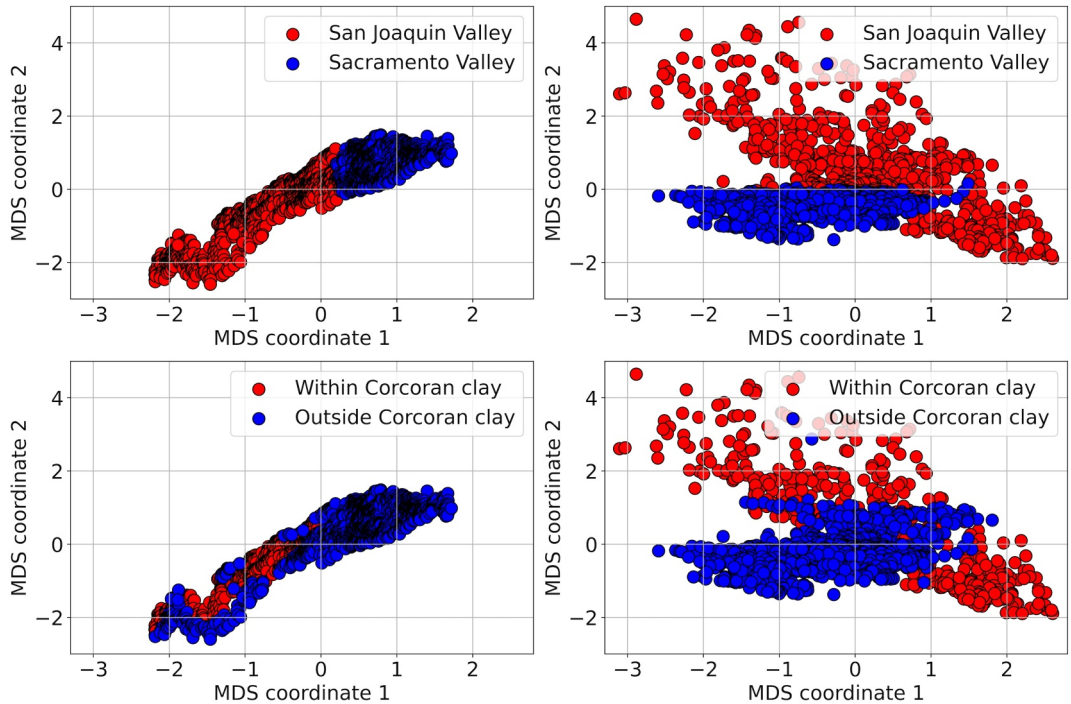

Figure 14. Training samples in compressed dimensions after MDS from $\boldsymbol{x}$-space (left) and $\tilde{\boldsymbol{x}}$-space (right).

observations generally indicate that the latent space reconfiguration was optimized to differentiate between locations based on their underlying lithological texture.

It is impossible to fully visualize and interpret how the latent space re-configuration varies with the texture properties in a 12-dimensional output space. Hence, we employ dimension reduction of the latent feature space to facilitate visual analysis. Specifically, we employ multidimensional scaling (MDS; Borg & Groenen, 1997), which projects samples from an input feature space into a lower-dimensional space while ensuring that the mutual distances between samples in the original space are preserved. This property of MDS is especially desirable since the covariance kernel is specified as a function of $\|\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{x}}'\|_2$. MDS has been employed for visualization of spatial subsurface uncertainty in several previous studies (Pradhan & Mukerji, 2020a, 2022; Scheidt & Caers, 2009). We compress the sample locations from the $\tilde{\boldsymbol{x}}$-space into a two dimensional space by metric MDS. The distance measure between samples $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{x}}'$ is taken to be $\|\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{x}}'\|_2$ which will be preserved during dimension reduction. Figure 14 compares the mutual distances between the training well locations in the $\boldsymbol{x}$-space against the $\tilde{\boldsymbol{x}}$-space after MDS. To support visual analysis, the $\tilde{\boldsymbol{x}}$-MDS and $\boldsymbol{x}$-MDS coordinates were separately normalized to have zero mean and unit variance. Note that the relative configuration of the samples along the MDS coordinates is representative with high accuracy of the relative configuration in the latent space. The correlation coefficient between Euclidean distances in the input feature space and Euclidean distances in the MDS space was calculated to be nearly 98%, indicating minimal loss of information during MDS. In subsequent discussion, references to the input feature space and its corresponding MDS space are used interchangeably.

In Figure 14, the samples have been colored by their native hydrological valley and whether the Corocoran Clay exists in the subsurface. It is immediately apparent that the DNN has learned to nicely separate based on the underlying aquifer textures. In the geospatial domain, locations without the clay confining unit exist both in the SV and eastern SJV. In the latent space, these sites have been gathered into a tight cluster. This explains the corresponding tighter confidence intervals observed in Figure 10 (compare top longitude transect against the bottom). Locations overlying the confining clay unit have been reconfigured with a distinct quasi-linear trend with a northwest to southeast orientation, with the semi-confined fine grained sediment equivalent thickness exhibiting a smooth gradation along this trend. The corresponding coarse grained sediment thicknesses also exhibit a clear southwest-northeast trending gradation (Figure 15), correlated with the scatter about the trend. Note especially that the variance about the northwest to southeast quasi-linear trend increases with in correlation

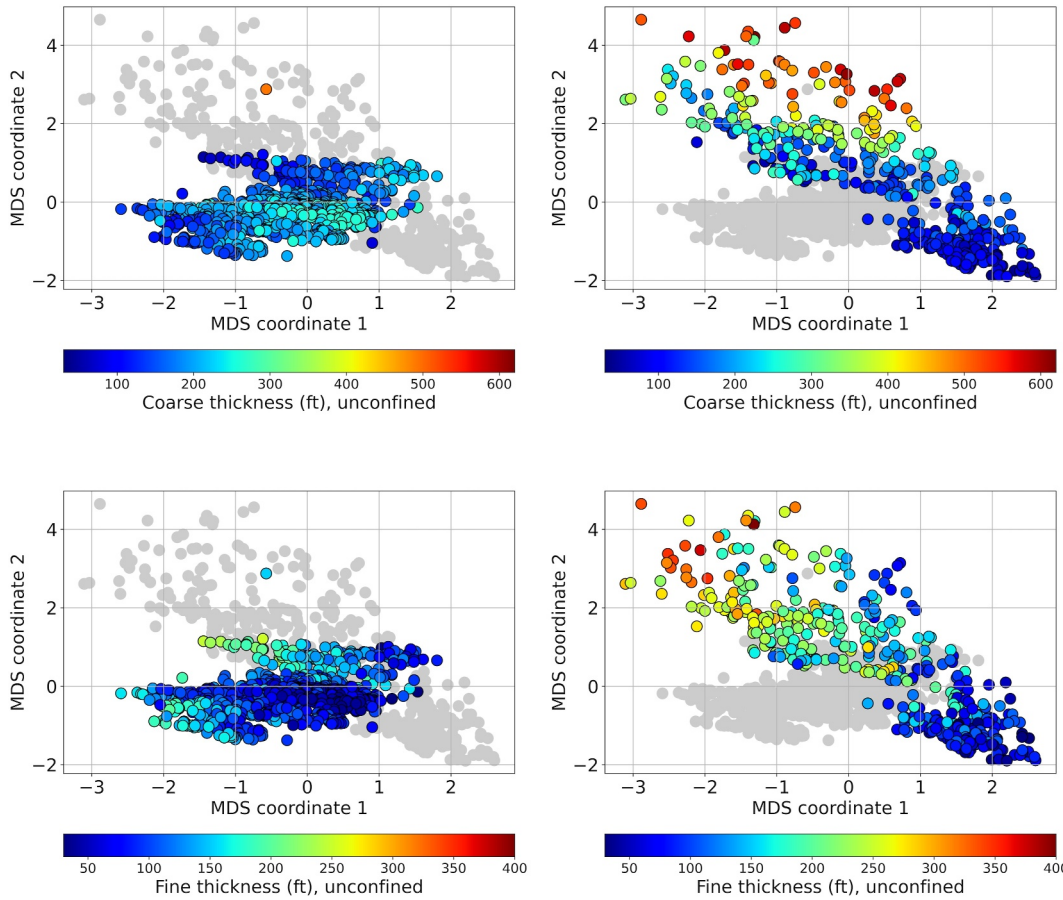

Figure 15. Training wells in $\tilde{\boldsymbol{x}}$-MDS space colored by thickness of coarse-grained sediments (top row) and fine-grained sediments (bottom row) in the shallow semi-confined aquifer zone. Locations with and without the Corcoran clay have been grayed out in the left and right columns respectively.

with fine grained (clayey) sediment thickness. This shows that the DNN has learned to differentiate between locations based on their underlying equivalent coarse and fine grained sediment column heights. From first principles, this is a natural expectation since two 1D aquifer columns with equal height, homogeneous sediment type and properties, and same boundary conditions will result in equivalent pressure head profiles (Equation 16). Greater the proportion of clay sediments in the aquifer column, the more heterogeneous the head profiles are expected to be. Thus, the flexibility of the GP-DNN formulation to distend and squash spatial distances facilitates it to (a) yield uncertainty predictions that are *texturally aware* instead of being just *geospatially aware*, as is typical in kriging-type models, and (b) model non-stationary spatial heterogeneity across the modeling domain. As validated with the blind well test set in Section 3.3, this leads to reliable uncertainty quantification.

4. Discussion

In this section, we discuss limitations and advantages of the GP-DNN regression model along with directions for future research. It should be noted that the CV texture model was obtained by non-stationary kriging of well data and has uncertainty associated with it. This is especially true along the sub-domain boundaries (for instance mapped edge of the Corcoran clay) and deeper aquifer where the well texture data is scant leading to artifacts in the model (Faunt, 2009). While we did not explicitly account for this uncertainty, we found empirically that the DNN ignores noisy features and artifacts especially from the deeper sections of the texture model. To improve reliability of the model predictions, additional robust features may be considered in the future. Such features could include (a) hydrogeophysical data such as electromagnetic data (Kang et al., 2021) which may provide spatially continuous information on aquifer structure and heterogeneity, (b) remote-sensing data such as InSAR surface deformation data which could inform on the sediment elastic/inelastic properties, (c) precipitation data, surface

water delivery data, and crop water use data which serve as proxies for groundwater source and sink flux. With additional data, it might be necessary to consider other regression models that can handle different data structures. For example, convolutional neural networks (CNNs; Krizhevsky et al., 2012) may be more suitable to learn from spatial geophysical data compared to DNNs (Pradhan & Mukerji, 2020b, 2022). The approach of hierarchically combining a regression model with GPs may in theory be extended to CNNs, with future effort required for applicable model design.

As discussed earlier, 3D variability of water levels resulting from vertical connectivity of aquifer layers was not considered and is a limitation of the current study. The proposed GP-DNN methodology may be extended to account for 3D effects, especially if well-screen depths may be combined with water level data to map the well observations as a function of latitude, longitude and depth. Notwithstanding the above limitation, one of the primary advantages of the GP-DNN model is the fast and analytical derivation of a statistically consistent posterior uncertainty model on long-term and seasonal groundwater trends informed by the lithological texture of the underlying aquifer layers. Informative probability distributions on key decision variables are a staple component for aiding decision making under uncertainty (Caers, 2011; Eidsvik et al., 2015). While the application was demonstrated specifically for the CV, similar data and uncertainty challenges exist in other hydrological basins where it should be possible to leverage latent space GPs for aiding uncertainty quantification.

It was shown earlier that the GP-DNN model did not significantly reduce the prior predictive uncertainty in the southern SJV. We identified two contributing factors that led to uninformative uncertainty quantification in this region: (a) highly variable hydrogeological heterogeneity, and (b) sparse well data. In the absence of sufficient information, the GP-DNN model correctly indicated that predictions were not reliable by putting large error bars on the posterior predictive mean in the southern SJV (Figure 10). A specific novelty of the GP-DNN model is that the uncertainty predictions are driven by the hydrogeological heterogeneity, in addition to spatial proximity of data observations, as compared to the traditional kriging-based approaches which only account for the latter. This resulted in predictions of tight uncertainty intervals even with sparse well data in certain regions, for instance compare data density in central to southern region of Solano subbasin (Figure 1) with the predicted uncertainty intervals (Figure 10).

A limitation with GP-DNN posterior predictive samples (Figure 9) is the apparent loss of spatial continuity in the southern SJV, with large swings of the variables observed between spatially close locations. In addition to gathering more data which will constrain the posterior predictive uncertainty as discussed above, we propose two future improvement directions specific to the methodology. In the proposed GP-DNN formulation, Gaussian smoothing was performed in the latent space and not directly on the spatial coordinates. Potential loss of spatial continuity may be prevented by enforcing spatial regularization directly. A simple trick to achieve this will be to augment the latent space with spatial information. Specifically, the latent space may be specified as $\tilde{\boldsymbol{x}} = \phi\left(\check{\boldsymbol{x}};\boldsymbol{\theta}\right)$, with the GP regression subsequently conducted in the augmented latent space $[\boldsymbol{x}, \tilde{\boldsymbol{x}}]^T$. In this case, the kernel length scales along the spatial and latent coordinates will need to be carefully tuned. The GP kernel will be expected to suppress large deviations of water levels within the specified spatial length-scales.

While the GP-DNN model allows accounting for observational noise, it was assumed that the noise level is spatially homogeneous across the valley. This assumption may be violated in southern SJV given that the density of the well data samples is heavily skewed toward the northern part of the valley. This could potentially lead to higher levels of noise during fitting of the trend parameters in Equation 15. This limitation may be addressed by specifying the noise level to vary spatially with each location. Since the true noise level is unknown, the noise level will have to be considered as a parameter of the model. Within the formulation presented in Section 2, it should be possible to derive the gradients of the loss function with respect to the noise parameters and train them end-to-end along with the neural network parameters. Spatially varying noise levels will provide the DNN flexibility to locally adjust the distances between the training locations in the latent space and prevent overfitting to any short correlations exhibited in the southern SJV data.

As discussed in Section 3.4.3, a common challenge associated with deep learning models is the difficulty associated in physically understanding how the input features affect the model outputs. In this paper, we approached model explainability by dimension reduction of the DNN estimated latent space and visualization of lithogical features along the latent reduced dimensions. This was effective in understanding how the DNN was effective in clustering together spatial locations with similar lithological characteristics. The reader is referred to

the review paper by Samek et al. (2021) for an overview of other state-of-the-art methods available for explaining deep learning models such as interpretable local surrogate models, feature perturbation methods, and layer-wise relevance propagation. A general limitation of covariance based approaches is that they scale as $\mathcal{O}(n^3)$ with the number of training samples n. While this was not a computational bottleneck for our training data set of 1,550 wells, this might introduce significant computational burden for larger data sets. Addressing computational challenges of covariance matrix based algorithms is a well studied research area. Rasmussen and Williams (2006) [Chapter 8] review several approximations methods that may be employed for applying GP-DNN to larger data sets.

5. Conclusions

A spatially continuous map of the groundwater level in CV is difficult to obtain due to the poor-quality of well data. In this paper, we proposed regression of sparse and noisy well data on features from a 3D lithological texture model of the CV aquifer system. We formulated a novel multivariate regression methodology that hierarchically leverages deep neural networks to morph the texture feature space into a latent space and Gaussian processes for non-parametric regression in the latent space. The proposed GP-DNN model provides a robust extension to traditional cokriging approach for modeling non-stationary data and augmenting uncertainty quantification with information from lithological features. We found that the GP-DNN model successfully captures non-stationary effects in the data by distending and squashing input distances in the latent space. The DNN was shown to be able to extract hydrogeologically explainable features from the data and the predictive uncertainty model was cross-validated to be statistically consistent with the empirical data distribution of 90 blind wells. Our results indicate that during 2015–2020 water levels in CV did not show appreciable recovery from the 2012–2015 drought in California. While the 2017 and 2019 wet years resulted in small and localized recovery of water levels, groundwater levels in August 2020 stayed mostly low in many areas of the valley. These results demonstrate promising applications of latent-space GP models to overcome data limitation challenges within hydrology and also have implications for refining our understanding of hydrologic connectivity in the context of groundwater recharge and drought recovery.

Appendix A: Construction of Covariance Matrices

The general anisotropic version of the stationary covariance kernel (Equation 7) is given as

$$k\left(\boldsymbol{x}, \boldsymbol{x}'\right) = \sum_i K^i_{amp} k^i_{valid}\left(\sqrt{(\boldsymbol{x}-\boldsymbol{x}')^T L^{-1}\left(\boldsymbol{x}-\boldsymbol{x}'\right)}\right)$$

where $\boldsymbol{x} \in \mathbb{R}^n$ and $L = \text{diag}\left(\left[l_1^2, l_2^2, \ldots, l_n^2\right]^T\right)$ is a diagonal matrix of the anisotropic length-scales along the input coordinates. The kernel length-scale may be related to the more traditional semivariogram range r, which is defined as the distance at which the semivariogram value reaches 95% of the semivariogram sill value (Goovaerts, 1997) as $l = \frac{r}{3}$. Note that the length-scales need to be specified for the baseline GP regression case. For simplicity, we assumed unit length-scales in the GP-DNN regression as the DNN can be expected to implicitly learn the scaling as part of the transformation $\phi(.)$.

The Matérn kernel used in Equation 8 is given as

$$k^{\nu}_{Matérn}(d) = \frac{2^{1-\nu}}{\Gamma(\nu)}(\sqrt{2\nu}\, d)^{\nu} K_{\nu}(\sqrt{2\nu}\, d),$$

where d is the scaled distance between two locations under evaluation, K_ν is a modified Bessel function, and ν is a roughness parameter. As a result of the linear model of coregionalization, the amplitude matrix K_{amp} is required to be positive semi-definite. K_{amp} controls the variance of the components of the multivariate random process and their correlation coefficients. We assumed each individual process to have unit variance and normalized the well training data appropriately. The off-diagonal elements of K_{amp} control the correlation coefficient between the processes since

$$\rho_{a_i(\boldsymbol{x}),\, a_j(\boldsymbol{x})} = \frac{cov[a_i(\boldsymbol{x}),\, a_j(\boldsymbol{x})]}{\sigma_{a_i(\boldsymbol{x})}\sigma_{a_j(\boldsymbol{x})}} = K_{amp}^{i,\,j} k_{Matérn}^{\nu=2.5}(\boldsymbol{x},\, \boldsymbol{x}) = K_{amp}^{i,\,j},$$

where ρ denotes the correlation coefficient, $cov[.]$ is the covariance operator and σ denotes standard deviation. Given training data samples $\{(\boldsymbol{x}_{\tau_i},\, \boldsymbol{y}_{\tau_1});\, i = 1,\, \ldots,\, m\}$ such that $\boldsymbol{x}_\tau \in \mathbb{R}^n$ and $\boldsymbol{y}_\tau \in \mathbb{R}^d$, the $4m \times 4m$ covariance matrix

$$K_{\tau\tau} = \begin{bmatrix} K_{\tau\tau}^{1,\,1} & \ldots & K_{\tau\tau}^{1,\,4} \\ \vdots & \ddots & \vdots \\ K_{\tau\tau}^{4,\,1} & \ldots & K_{\tau\tau}^{4,\,4} \end{bmatrix}, \tag{A1}$$

where,

$$K_{\tau\tau}^{i,\,j} = \begin{bmatrix} k^{i,\,j}(\boldsymbol{x}_{\tau_1},\, \boldsymbol{x}_{\tau_1}) & \ldots & k^{i,\,j}(\boldsymbol{x}_{\tau_1},\, \boldsymbol{x}_{\tau_m}) \\ \vdots & \ddots & \vdots \\ k^{i,\,j}(\boldsymbol{x}_{\tau_m},\, \boldsymbol{x}_{\tau_1}) & \ldots & k^{i,\,j}(\boldsymbol{x}_{\tau_m},\, \boldsymbol{x}_{\tau_m}) \end{bmatrix}, \quad \forall\, i, j = 1,\, \ldots,\, d.$$

In the above,

$$k^{i,\,j}(\boldsymbol{x}_{\tau_k},\, \boldsymbol{x}_{\tau_l}) = K_{amp}^{i,\,j} k_{Matérn}^{\nu=2.5}(\boldsymbol{x}_{\tau_k},\, \boldsymbol{x}_{\tau_l}).$$

Appendix B: Time Series Linear Regression

The mathematical model capturing long-term and seasonal trends in well water level time series data is given in Equation 15. We derive how the parameters c may be estimated independently at each training location by linear regression. Equation 15 can be re-written as

$$u(\boldsymbol{x},\, t) = a_1(\boldsymbol{x}) + a_2(\boldsymbol{x})t + a_3'(\boldsymbol{x})\sin\left(\frac{2\pi t}{\lambda}\right) + a_4'(\boldsymbol{x})\cos\left(\frac{2\pi t}{\lambda}\right), \tag{B1}$$

where, $a_3'(\boldsymbol{x}) = a_3(\boldsymbol{x})\cos(a_4(\boldsymbol{x}))$ and $a_4'(\boldsymbol{x}) = a_3(\boldsymbol{x})\sin(a_4(\boldsymbol{x}))$. Given data $\{(t_j,\, u_{\boldsymbol{x}_{\tau_i},\, j});\, j = 1,\, \ldots,\, n_i\}$ at the i^{th} well, Equation B1 may be used to specify the following system of linear equations

$$\boldsymbol{u}_{\tau_i} = X_{\tau_i}\boldsymbol{a}_{\tau_i}, \tag{B2}$$

where, $\boldsymbol{u}_{\tau_i}$ is the $n_i \times 1$ vector of water levels, X_{τ_i} is the $n_i \times 4$ feature matrix and $\boldsymbol{a}_{\tau_i} = [a_1(\boldsymbol{x}_{\tau_i}),\, a_2(\boldsymbol{x}_{\tau_i}),\, a_3'(\boldsymbol{x}_{\tau_i}),\, a_4'(\boldsymbol{x}_{\tau_i})]^T$ is the trend parameter vector to be estimated. The least-squares solution to Equation B2 is given as

$$\boldsymbol{a}_{\tau_i} = \left(X_{\tau_i}^T X_{\tau_i}\right)^{-1} X_{\tau_i}^T \boldsymbol{u}_{\tau_i}. \tag{B3}$$

We solve Equation B3 independently at each well location to obtain estimates of the corresponding parameter vector $\boldsymbol{a}_{\tau_i}$. The seasonal amplitude a_3 and phase delay a_4, computed as

$$a_3 = \sqrt{a_3'^2 + a_4'^2} \text{ and } a_4 = \arctan 2\left(a_3',\, a_4'\right)$$

respectively.

Appendix C: Training and Tuning of Regression Models

C1. Baseline GP Regression

In this case, the regression is performed in the geospatial coordinate space $\boldsymbol{x}$. The posterior predictive distribution may be analytically derived as shown in Equation 6. The list of hyper-parameters is shown in Table C1. For each hyper-parameter, we specify a range of possible values it may assume. In general, available data at training wells were used as a guide to specify the support of the hyper-parameters. For instance, to estimate the ranges for length-scales l_1 and l_2 along latitude and longitude coordinates, empirical variogram analysis (Goovaerts, 1997) with well data indicated that length-scales ranged roughly between 5 and 15 miles. To account for the uncertainty in the empirical estimates, we considered lower and upper bounds of 3 and 75 miles respectively. The other hyper-parameters considered are related to the covariance kernel specification, that is, K_{amp} and Σ_n. The diagonal elements of the Matérn kernel amplitude matrix K_{amp}, defining the signal variances of a_i, $i = 1, \ldots, 4$, are taken to be 1 given that we normalized all the training, validation and test data to have unit variance. Uncertainty in the diagonal entries of the covariance matrix is modeled through the Σ_n described below. The off-diagonal entries of the $K_{amp}^{i,j}$, $i \neq j$, capture the correlation coefficients between the trend parameters (Appendix A). Note that there are only six free off-diagonal elements, since K_{amp} is required to be positive semi-definite by definition. We consider higher noise variances for a_3 and a_4 since estimating the sinusoidal phase parameter a_4' from the sparse well time series will typically be more difficult than intercept, slope and sinusoidal amplitude, resulting in noisier estimates of a_3 and a_4 from Equation B3. Hyper-parameter values tuned by cross-validation are shown in Table C1.

Table C1
Hyper-Parameter Tuning Details for Baseline GP Regression

Hyper-parameter	Parameter range	Tuned value
Length-scale l_1	[3 miles, 75 miles]	12.13 miles
Length-scale l_2	[3 miles, 75 miles]	7.11 miles
Noise variance $\sigma_{n_1}^2$	[0.2, 0.5]	0.23
Noise variance of $\sigma_{n_2}^2$	[0.2, 0.7]	0.61
Noise variance of $\sigma_{n_3}^2$	[0.4, 0.98]	0.43
Noise variance of $\sigma_{n_4}^2$	[0.4, 0.98]	0.70
Kernel amplitude $K_{amp}^{1,2}$	[−0.7, −0.25]	−0.41
Kernel amplitude $K_{amp}^{1,3}$	[0.01, 0.2]	0.16
Kernel amplitude $K_{amp}^{1,4}$	[−0.4, −0.05]	−0.37
Kernel amplitude $K_{amp}^{2,3}$	[−0.15, 0.15]	0.10
Kernel amplitude $K_{amp}^{2,4}$	[−0.05, 0.15]	−0.02
Kernel amplitude $K_{amp}^{3,4}$	[−0.9, −0.4]	−0.80

C2. Hierarchical GP-DNN Regression

In this case, the prior predictive distribution is specified using the hierarchical model posited in Equations 2 and 4. The DNN in the bottom layer may be parameterized through several hidden layers, each of which constitutes of a number of neurons with trainable weight and bias parameters $\boldsymbol{\theta}$ to yield multivariate outputs. We treat the number of hidden layers, neurons and outputs as hyper-parameters taking values in the ranges shown in Table C2. At the top-level is a GP model, regressed in the space of DNN outputs $\tilde{\boldsymbol{x}}$. The posterior predictive distribution on prediction variables may then be derived as shown in Equation 12. Given above model specification, parameters $\boldsymbol{\theta}$ will be optimized by minimizing the negative log-likelihood of the GP marginal distribution, and hyper-parameters tuned by cross validation with the negative log-likelihood of the validation data under the GP posterior predictive distribution. To limit overfitting, we considered dropout (Srivastava et al., 2014) and ℓ_2-regularization, with dropout regularization found to be largely ineffective (Table C2). Based on cross-validation, the optimal DNN architecture was found to consist of two hidden layers with 33 neurons in each layer. The dimension

Table C2
Hyper-Parameter Tuning Details for GP-DNN Regression

Hyper-parameter	Parameter range	Tuned value
Number of hidden layers	{1, 2, 3}	2
Number of neurons in hidden layers	{30, 31, …, 130}	33
Number of DNN output nodes	{1, 2, …, 30}	12
Use dropout	{True, False}	False
Learning rate for DNN training with Adam optimization algorithm	[0.001, 0.5]	0.28
Weight of ℓ_2-regularization	[0.001, 10]	2.91
Noise variance $\sigma^2_{n_1}$	[0.2, 0.5]	0.28
Noise variance of $\sigma^2_{n_2}$	[0.2, 0.7]	0.54
Noise variance of $\sigma^2_{n_3}$	[0.4, 0.98]	0.93
Noise variance of $\sigma^2_{n_4}$	[0.4, 0.98]	0.85
Kernel amplitude $K^{1,\,2}_{amp}$	[−0.7, −0.25]	−0.42
Kernel amplitude $K^{1,\,3}_{amp}$	[0.01, 0.2]	0.11
Kernel amplitude $K^{1,\,4}_{amp}$	[−0.4, −0.05]	−0.28
Kernel amplitude $K^{2,\,3}_{amp}$	[−0.05, 0.15]	−0.06
Kernel amplitude $K^{2,\,4}_{amp}$	[−0.05, 0.15]	0.03
Kernel amplitude $K^{3,\,4}_{amp}$	[−0.9, −0.4]	−0.47

of the latent space was tuned to be 12. In addition to the DNN hyper-parameters, we also consider hyper-parameters related to the GP Mátern kernel and data noise. For simplicity, we assumed unit length-scales $\boldsymbol{l} \in \mathbb{R}^{p\times 1}$.

The training procedure is shown in Algorithm 1. The training algorithm takes as inputs the training data feature matrix $\hat{X}_\tau \in \mathbb{R}^{m\times(n+2)}$, where $m = 1550$ is the number of training examples and $n = 39$ is the dimension of the hydrogeological feature space, and the target vector $\boldsymbol{y}_\tau \in \mathbb{R}^{4m\times 1}$. In Figure C1, we show the variability of the training set loss function across 100 training epochs. Note that the loss function in Equation 10 comprises of the data fit and model complexity terms (see similar discussion for the posterior predictive distribution in Section 2.2.4.1). As the training progresses, the DNN model parameters $\boldsymbol{\theta}$ are expected to become increasingly complex to overfit to the training data leading to larger values of the model complexity term. This is clearly observed in the increasing trend of the loss function beyond the tenth epoch, driven by increasing model complexity term. To ensure that the DNN model parameters may generalize to data sets other than the training set, we enforce early stopping of the training based on the cross-validation metric of negative log-likelihood of the predictive posterior distribution evaluated on the validation set. In the right subplot of Figure C1, we show the behavior of validation set cross-validation metric across 100 training epochs. The best validation metric is attained at the seventh epoch, hence the DNN parameters $\boldsymbol{\theta}$ from this epoch are employed for all subsequent model predictions.

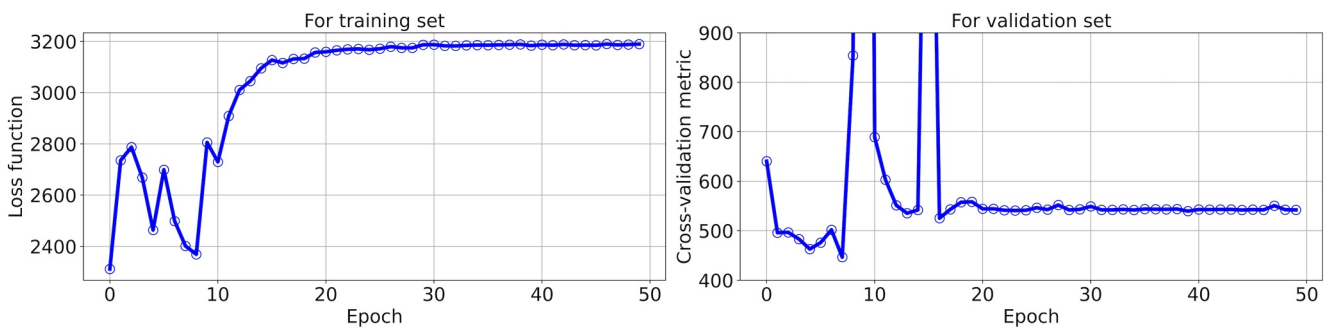

Figure C1. GP-DNN training history across 100 epochs. (left) Negative of marginal log-likelihood loss function (R.H.S. of Equation 10 without the constant term) evaluated on the training set is plotted along the *y*-axis. (right) Negative of predictive posterior log-likelihood evaluated on the validation set is plotted along the *y*-axis.

Data Availability Statement

The CV lithologic texture data is available via the United States Geological Survey data release at https://doi.org/10.5066/P9IZRO3V (Marcelli et al., 2022), and the CV digital elevation model is available at https://doi.org/10.5067/MEaSUREs/NASADEM/NASADEM_HGT.001 (NASA JPL, 2020). The CV well water level data is attributed to Kim et al. (2021). The well data, processed as described in Section 3.1 along with the GP-DNN modeled water level trends and time series outputs, may be accessed through the Harvard Dataverse repository at https://doi.org/10.7910/DVN/23TNJO with Creative Commons Attribution 4.0 International license (Pradhan et al., 2024). Version v0.1.0 of Python software for GP-DNN regression, written using open source Tensorflow (TensorFlow Developers, 2023), NumPy (Harris et al., 2020) and Scipy (Virtanen et al., 2020) libraries, is preserved at https://doi.org/10.5281/zenodo.13855361, available via Creative Commons Attribution 4.0 International license (Pradhan, 2024). Normal score transformation method was performed using mGstat geostatistical MATLAB toolbox (Hansen, 2022). All data analyses were performed using open source NumPy and Scipy Python libraries, while data visualizations were conducted using open source Matplotlib Python library (Caswell et al., 2021; Hunter, 2007) and its Basemap extension.

Acknowledgments
We thank Caltech's Resnick Sustainability Institute for funding the work presented and Professor Mark Simons and Dr. Neil Fromer for helpful discussions on groundwater modeling. Professor Venkat Chandrasekaran was supported in part by Air Force Office of Scientific Research (AFOSR) Grants FA9550-23-1-0204 and FA9550-22-1-0225, and by NSF Grant DMS 2113724. Professor Andrew M. Stuart gratefully acknowledges support by the AFOSR under MURI award number FA9550-20-1-0358 (Machine Learning and Physics-Based Modeling and Simulation). We offer thanks to Dr. Kyongsik Yun and Dillon Holder for help with the processing of the data shown. We are also grateful towards Professor Tapan Mukerji whose expertise with geostatistics helped provide valuable insights supporting the presented research.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international Conference on Knowledge Discovery & Data Mining* (pp. 2623–2631). Association for Computing Machinery. https://doi.org/10.1145/3292500.3330701

Bertoldi, G. L., Johnston, R. H., & Evenson, K. D. (1991). *Ground water in the central valley, California—A summary report* (Vol. 1401-A). U.S. Geological Survey Professional Paper.

Bertoncello, A. (2011). *Conditioning surface-based models to well and thickness data* (Doctoral Dissertation). Stanford University. Retrieved from https://stacks.stanford.edu/file/druid:tr997gp6153/DISSERTATION_Antoine_Bertoncello-augmented.pdf

Binley, A., Hubbard, S. S., Huisman, J. A., Revil, A., Robinson, D. A., Singha, K., & Slater, L. D. (2015). The emergence of hydrogeophysics for improved understanding of subsurface processes over multiple scales. *Water Resources Research*, *51*(6), 3837–3866. https://doi.org/10.1002/2015WR017016

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer-Verlag.

Borg, I., & Groenen, P. (1997). *Modern multidimensional scaling: Theory and applications*. Springer.

Bradshaw, J., Matthews, A. G. d. G., & Ghahramani, Z. (2017). Adversarial examples, uncertainty, and transfer testing robustness in Gaussian process hybrid deep networks. *arXiv*. https://doi.org/10.48550/ARXIV.1707.02476

Caers, J. (2005). *Petroleum geostatistics*. Society of Petroleum Engineers.

Caers, J. (2011). *Modeling uncertainty in the earth sciences*. John Wiley and Sons.

Calandra, R., Peters, J., Rasmussen, C. E., & Deisenroth, M. P. (2016). Manifold Gaussian processes for regression. *International Joint Conference on Neural Networks*, *51*, 3338–3345.

California Department of Water Resources. (2003). *California's ground water, update 2003* (Tech. Rep. No. 118). California Department of Water Resources.

California Department of Water Resources. (2017). *Fall 2017 groundwater level data summary* (Tech. Rep.). California Department of Water Resources. Retrieved from https://water.ca.gov/-/media/DWR-Website/Web-Pages/Programs/Groundwater-Management/Data-and-Tools/Files/Statewide-Reports/Fall-2017-Groundwater-Level-Data-Summary_ay_19.pdf

California Department of Water Resources. (2019). *California groundwater conditions update—Spring 2019* (Tech. Rep.). California Department of Water Resources. Retrieved from https://water.ca.gov/-/media/DWR-Website/Web-Pages/Programs/Groundwater-Management/Data-and-Tools/Files/Maps/Groundwater-Level-Change/DOTMAP_Reports/Spring-2019-Groundwater-DOTMAP-Report.pdf

Caswell, T. A., Droettboom, M., Lee, A., de Andrade, E. S., Hoffmann, T., Hunter, J., et al. (2021). matplotlib/matplotlib: Rel: v3.5.1 [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.5773480

Chaussard, E., Bürgmann, R., Shirzaei, M., Fielding, E. J., & Baker, B. (2014). Predictability of hydraulic head changes and characterization of aquifer-system and fault properties from InSAR-derived ground deformation. *Journal of Geophysical Research: Solid Earth*, *119*(8), 6572–6590. https://doi.org/10.1002/2014JB011266

Cressie, N. (1993). *Statistics for spatial data*. Wiley.

Damianou, A. C., & Lawrence, N. D. (2013). Deep Gaussian processes. *Artificial Intelligence and Statistics*, *31*, 207–215.

De Iaco, S., Myers, D., & Posa, D. (2003). The linear coregionalization model and the product-sum space-time variogram. *Mathematical Geology*, *35*(1), 25–38. https://doi.org/10.1023/A:1022425111459

de Marsily, G. (1987). Non stationary geostatistics. In J. Bear & M. Y. Corapcioglu (Eds.), *Advances in transport phenomena in porous media* (Vol. 128, pp. 633–655). Springer. https://doi.org/10.1007/978-94-009-3625-6_13

Deutsch, C. V., & Journel, A. G. (1998). *Gslib: Geostatistical software library and user's guide*. Oxford University Press.

Dunlop, M. M., Girolami, M. A., Stuart, A. M., & Teckentrup, A. L. (2018). How deep are deep Gaussian processes? *Journal of Machine Learning Research*, *19*(1), 2100–2145.

Easton, G. S., & McCulloch, R. E. (1990). A multivariate generalization of quantile-quantile plots. *Journal of the American Statistical Association*, *85*(410), 376–386. https://doi.org/10.1080/01621459.1990.10476210

Eidsvik, J., Mukerji, T., & Bhattacharjya, D. (2015). *Value of information in the earth sciences: Integrating spatial modeling and decision analysis*. Cambridge University Press.

Etherington, T. R. (2019). Mahalanobis distances and ecological niche modelling: Correcting a chi-squared probability error. *PeerJ*, *7*, e6678. https://doi.org/10.7717/peerj.6678/supp-1

Faunt, C. C. (2009). *Groundwater availability of the central valley aquifer, California (No. 1766)*. U.S. Geological Survey Professional Paper.

Faunt, C. C., Sneed, M., Traum, J., & Brandt, J. T. (2016). Water availability and land subsidence in the central valley, California, USA. *Hydrogeology Journal*, *24*(3), 675–684. https://doi.org/10.1007/s10040-015-1339-x

Feyen, L., & Caers, J. (2006). Quantifying geological uncertainty for flow and transport modeling in multi-modal heterogeneous formations. *Advances in Water Resources*, *29*(6), 912–929. https://doi.org/10.1016/j.advwatres.2005.08.002

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (Vol. 9, pp. 249–256).

Gnanadesikan, R., & Wilk, M. B. (1968). Probability plotting methods for the analysis of data. *Biometrika*, *55*, 1–17. https://doi.org/10.2307/2334448

Goovaerts, P. (1997). *Geostatistics for natural resources evaluation*. Oxford University Press.

Gualandi, A., & Liu, Z. (2021). Variational Bayesian independent component analysis for InSAR displacement time-series with application to central California, USA. *Journal of Geophysical Research: Solid Earth*, *126*(4), e2020JB020845. https://doi.org/10.1029/2020JB020845

Hansen, T. M. (2022). mgstat [Software]. https://github.com/cultpenguin/mGstat/releases/tag/1.1

Harbaugh, A. W. (2005). *Modflow-2005: The U.S. geological survey modular ground-water model–the ground-water flow process* (Tech. Rep. No. 6-A16). USGS. https://doi.org/10.3133/tm6A16

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hermans, T., Nguyen, F., & Caers, J. (2015). Uncertainty in training imagebased inversion of hydraulic head data constrained to ERT data: Workflow and case study. *Water Resources Research*, *51*(7), 5332–5352. https://doi.org/10.1002/2014WR016460

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science and Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Johnson, D., & Belitz, K. (2015). Identifying the location and population served by domestic wells in California. *Journal of Hydrology*, *3*, 31–86. https://doi.org/10.1016/j.ejrh.2014.09.002

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis*. Pearson.

Journel, A. G., & Huijbregts, C. J. (1978). *Mining geostatistics*. Academic Press.

Kang, S., Knight, R., Greene, T., Buck, C., & Fogg, G. (2021). Exploring the model space of airborne electromagnetic data to delineate large scale structure and heterogeneity within an aquifer system. *Water Resources Research*, *57*(10), e2021WR029699. https://doi.org/10.1029/2021WR029699

Keating, E. H., Doherty, J., Vrugt, J. A., & Kang, Q. (2013). Optimization and uncertainty assessment of strongly nonlinear groundwater models with high parameter dimensionality. *Water Resources Research*, *46*(10), W10517. https://doi.org/10.1029/2009WR008584

Kim, K. H., Liu, Z., Rodell, M., Beaudoing, H., Massoud, E., Kitchens, J., et al. (2021). An evaluation of remotely sensed and in situ data sufficiency for SGMA-scale groundwater studies in the central valley, California. *JAWRA Journal of the American Water Resources Association*, *57*(5), 664–674. https://doi.org/10.1111/1752-1688.12898

Kingma, D., & Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems in the witwatersand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, *52*, 119–139.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, & K. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 25). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Laloy, E., Rogiers, B., Vrugt, J. A., Mallants, D., & Jacques, D. (2013). Efficient posterior exploration of a high-dimensional groundwater model from two-stage Markov Chain Monte Carlo simulation and polynomial chaos expansion. *Water Resources Research*, *49*(5), 2664–2682. https://doi.org/10.1002/wrcr.20226

Laudon, J., & Belitz, K. (1991). Texture and depositional history of late pleistocene–holocene alluvium in the central part of the western San Joaquin Valley, California Tech. Rep. No. 1. *Bulletin of the Association of Engineering Geologists*, *28*.

Linde, N., Renard, P., Mukerji, T., & Caers, J. (2015). Geological realism in hydrogeological and geophysical inverse modeling: A review. *Advances in Water Resources*, *86*, 86–101. https://doi.org/10.1016/j.advwatres.2015.09.019

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B*, *73*(4), 423–498. https://doi.org/10.1111/j.1467-9868.2011.00777.x

Liu, P. W., Famiglietti, J. S., Purdy, A. J., Adams, K. H., McEvoy, A. L., Reager, J. T., et al. (2022). Groundwater depletion in California's central valley accelerates during megadrought. *Nature Communications*, *13*(1), 7825. https://doi.org/10.1038/s41467-022-35582-x

Liu, Z., Liu, P.-W., Massoud, E., Farr, T. G., Lundgren, P., & Famiglietti, J. S. (2019). Monitoring groundwater change in California's central valley using Sentinel-1 and grace observations. *Geosciences*, *9*(10), 436. https://doi.org/10.3390/geosciences9100436

Mahalanobis, P. C. (1936). On the generalised distance in statistics. In *Proceedings of the National Institute of Sciences of India* (Vol. 2, pp. 49–55).

Marcelli, M. F., Shepherd, M. M., & Faunt, C. C. (2022). Central valley hydrologic model version 2 (CVHM2): Well log lithology database and texture model [Dataset]. *U.S. Geological Survey*. https://doi.org/10.5066/P9IZRO3V

Mariethoz, G., & Caers, J. (2014). *Multiple point geostatistics: Stochastic modeling with training images*. Wiley Blackwell.

Mariethoz, G., Renard, P., & Caers, J. (2010). Bayesian inverse problem and optimization with iterative spatial resampling. *Water Resources Research*, *46*(11), W11530. https://doi.org/10.1029/2010WR009274

Matheron, G. (1962). *Trait'e de geostatistique appliquee*. Technip.

Mavko, G., T, M., & J, D. (2009). *The rock physics handbook*. Cambridge University Press.

NASA JPL. (2020). NASADEM merged DEM global 1 arc second v001 [Dataset]. *Distributed by NASA EOSDIS Land Processes DAAC*. https://doi.org/10.5067/MEaSUREs/NASADEM/NASADEM_HGT.001

Neely, W. R., Borsa, A. A., Burney, J. A., Levy, M. C., Silverii, F., & Sneed, M. (2021). Characterization of groundwater recharge and flow in California's San Joaquin Valley from InSAR-observed surface deformation. *Water Resources Research*, *57*(4), e2020WR028451. https://doi.org/10.1029/2020wr028451

Ojha, C., Werth, S., & Shirzaei, M. (2019). Groundwater loss and aquifer system compaction in San Joaquin valley during 2012–2015 drought. *Journal of Geophysical Research: Solid Earth*, *124*(3), 3127–3143. https://doi.org/10.1029/2018jb016083

Paciorek, C. J. (2003). *Nonstationary Gaussian processes for regression and spatial modelling* (Doctoral Dissertation). Carnegie Mellon University. Retrieved from https://www.stat.berkeley.edu/~paciorek/diss/paciorek-thesis.pdf

Pradhan, A. (2020). *Statistical learning and inference of subsurface properties under complex geological uncertainty with seismic data* (Doctoral Dissertation). Stanford University. Retrieved from https://stacks.stanford.edu/file/druid:sg756xs1514/Dissertation_AnshumanPradhan-augmented.pdf

Pradhan, A. (2024). pradhan-a/gp-dnn: Gp-dnn code release [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.13855361

Pradhan, A., Adams, K. H., Chandrasekaran, V., Liu, Z., Reager, J. T., Stuart, A. M., & Turmon, M. J. (2024). Modeled groundwater levels across Central Valley, CA, from March 2015 to August 2020, using GP-DNN regression [Dataset]. *Harvard Dataverse*. https://doi.org/10.7910/DVN/23TNJO

Pradhan, A., & Mukerji, T. (2020a). Seismic Bayesian evidential learning: Estimation and uncertainty quantification of sub-resolution reservoir properties. *Computational Geosciences*, *24*(3), 1121–1140. https://doi.org/10.1007/s10596-019-09929-1

Pradhan, A., & Mukerji, T. (2020b). Seismic inversion for reservoir facies under geologically realistic prior uncertainty with 3d convolutional neural networks. In *Seg technical program expanded abstracts 2020* (pp. 1516–1520). https://doi.org/10.1190/segam2020-3426944.1

Pradhan, A., & Mukerji, T. (2022). Consistency and prior falsification of training data in seismic deep learning: Application to offshore deltaic reservoir characterization. *Geophysics*, *87*(3), N45–N61. https://doi.org/10.1190/geo2021-0568.1

Pyrcz, M. J., & Deutsch, C. V. (2014). *Geostatistical reservoir modeling*. Oxford University Press.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.

Riel, B., Simons, M., Ponti, D., Agram, P., & Jolivet, R. (2018). Quantifying ground deformation in the Los Angeles and Santa Ana coastal basins due to groundwater withdrawal. *Water Resources Research*, *54*(5), 3557–3582. https://doi.org/10.1029/2017wr021978

Roininen, L., Girolami, M., Lasanen, S., & Markkanen, M. (2019). Hyperpriors for Matérn fields with applications in Bayesian inversion. *Inverse Problems and Imaging*, *13*, 1–29. https://doi.org/10.3934/ipi.2019001

Rousseeuw, P. J., & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, *41*(3), 212–223. https://doi.org/10.1080/00401706.1999.10485670

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K.-R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, *109*(3), 247–278. https://doi.org/10.1109/JPROC.2021.3060483

Sampson, P., Damian, D., & Guttorp, P. (2001). Advances in modeling and inference for environmental processes with nonstationary spatial covariance. In *GeoENV III: Geostatistics for Environmental Applications* (pp. 17–32).

Sampson, P. D., & Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, *87*(417), 108–119. https://doi.org/10.2307/2290458

Scheidt, C., & Caers, J. (2009). Representing spatial uncertainty using distances and kernels. *Mathematical Geosciences*, *41*(4), 397–419. https://doi.org/10.1007/s11004-008-9186-0

Schmidt, A. M., & O'Hagan, A. (2003). Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society: Series B*, *65*(3), 743–758. https://doi.org/10.1111/1467-9868.00413

Smith, R., & Knight, R. (2019). Modeling land subsidence using InSAR and airborne electromagnetic data. *Water Resources Research*, *55*(4), 2801–2819. https://doi.org/10.1029/2018WR024185

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958.

Stein, M. L. (1999). *Interpolation of spatial data*. Springer-Verlag.

Stuart, A. M. (2010). *Inverse problems: A Bayesian perspective*. Cambridge University Press.

Tarantola, A. (2005). *Inverse problem theory and methods for model parameter estimation*. SIAM.

TensorFlow Developers. (2023). Tensorflow [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.7604226

Todd, D., & Mays, L. (2005). *Groundwater hydrology*. John Wiley and Sons.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, *24*(3–4), 471–494. https://doi.org/10.1093/biomet/24.3-4.471

Williamson, A. K., Prudic, D. E., & Swain, L. A. (1989). In *Ground-water flow in the central valley, California* (Vol. 1401-D). U.S. Geological Survey Professional Paper.

Wilson, A. G., Hu, Z., Salakhutdinov, R., & Xing, E. P. (2016). Deep kernel learning. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics* (Vol. 51, pp. 370–378).